ELSEVIER

Speech Communication 33 (2001) 135–151

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Automatic ToBI prediction and alignment to speed manual labeling of prosody

Ann K. Syrdal [a,*], Julia Hirschberg [a], Julie McGory [b], Mary Beckman [b]

[a] *AT&T Labs. – Research, Florham Park, NJ, USA*

[b] *Department of Linguistics, Ohio State University, Columbus, OH, USA*

Accepted 2 August 2000

Abstract

Tagging of corpora for useful linguistic categories can be a time-consuming process, especially with linguistic categories for which annotation standards are relatively new, such as discourse segment boundaries or the intonational events marked in the Tones and Break Indices (ToBI) system for American English. A ToBI prosodic labeling of speech typically takes even experienced labelers from 100 to 200 times real time. An experiment was conducted to determine (1) whether manual correction of automatically assigned ToBI labels would speed labeling, and (2) whether default labels introduced any bias in label assignment. A large speech corpus of one female speaker reading several types of texts was automatically assigned default labels. Default accent placement and phrase boundary location were predicted from text using machine learning techniques. The most common ToBI labels were assigned to these locations for default tones and break type. Predicted pitch accents were automatically aligned to the mid-point of the word, while breaks and edge tones were aligned to the end of the phrase-final word. The corpus was then labeled by a group of five trained transcribers working over a period of nine months. Half of each set of recordings was labeled in the standard fashion without default labels, and the other half was presented with preassigned default labels for labelers to correct. Results indicate that labeling from defaults was generally faster than standard labeling, and that defaults had relatively little impact on label assignment.

1. Introduction

In the decade or so since the development of the TIMIT (Garofolo et al., 1986) and Penn TreeBank (Marcus et al., 1993) databases, the importance of large annotated language corpora has been firmly established. For example, the TIMIT database has been used in research on American English pronunciation styles (e.g., Byrd, 1992), and the consonant and vowel allophone labels for which it is tagged still serve as a standard industry-wide tool for training and testing automatic speech recognition (ASR) systems for the language. However, the tagging of such useful linguistic categories can be a very labor-intensive process. For example, phonetic segment tagging of a new speech corpus still takes at least an order of magnitude longer than the actual speech time even for standard American English, a language for which there is a good consensus about the relevant consonant and vowel categories based on decades of research on American pronunciation (see, e.g., Kenyon and Knott, 1953; Olive et al., 1993), well-established segmentation criteria dating back to the 1950s

* Corresponding author.

0167-6393/01/$ - see front matter
PII: S0167-6393(00)00073-X

(see, e.g., House and Fairbanks, 1953; Lehiste, 1960; Peterson and Lehiste, 1960), and the engineering of standard tools for automating some aspects of the labeling (e.g., Wightman and Talkin, 1994). The cost of tagging can seem especially prohibitive when a new tagging schema is first proposed – i.e., before there is a long enough history of use to develop firm standards for training and testing new taggers, and a large enough body of tagged data to use in establishing methods for automating some aspects of the tagging process.

In this paper, we describe an approach to enable faster tagging of spoken language corpora for prosodic labels. The labels are the American English ToBI categories for intonational events and degree of inter-word disjunctures, and the approach was inspired by a standard approach to increase the speed of segmental (phonetic) labeling. For segment labeling, the method is to use an online pronunciation dictionary coupled with ASR technology to automatically assign and align phone labels to a speech sample. Manual correction is then done by trained segment labelers, using an off-the-shelf spectrogram display cum waveform annotation utility. The analogous method for prosodic tags uses text-to-speech (TTS) synthesis technology to automatically assign and align prosodic labels from the transcribed text, before handing the labels over to a trained ToBI labeler for correction.

1.1. The ToBI system of prosodic labels

The tagging schema used to describe prosodic phenomena below is the ToBI model for standard American English (Pitrelli et al., 1994; Silverman et al., 1992). [1] The AmE–ToBI system consists of annotations at four or more time-linked levels of analysis. The three obligatory tiers are: an ORTHOGRAPHIC TIER of time-aligned words; a BREAK INDEX TIER indicating degrees of junction between words, from 0 'no word boundary' to 4 'full intonational phrase boundary' (Price et al., 1991); and a TONAL TIER, where PITCH ACCENTS, PHRASE ACCENTS and BOUNDARY TONES describing targets in the FUNDAMENTAL FREQUENCY (f_0) define intonational contours, roughly following Pierrehumbert's (Pierrehumbert and Hirschberg, 1990) scheme for American English. A fourth tier, the MISCELLANEOUS TIER, is provided for any additional phenomena that particular groups may wish to tag, such as dysfluencies. Other site-specific tiers are also encouraged, and later we describe a site-specific tier added for the purpose of marking points of discussion among the transcribers.

Break indices and tones together define two levels of phrasing, minor or intermediate (level 3) and major or intonational (level 4), with the former delimited by a phrase accent and the latter delimited by a sequence of phrase accent and boundary tone. That is, a level 3 phrase consists of a (string of) word(s) with at least one pitch accent aligned with the rhythmically strongest syllable of the accented lexical item(s), followed by a phrase accent which may be high (***H***–) or low (***L***–), and level 4 phrases consist of one or more level 3 phrases, plus a high or low boundary tone (***H***% or ***L***%) at the right edge of the phrase. A standard declarative pitch contour, for example, ends in a low phrase accent and low boundary tone, and is represented by ***L*** – ***L***%; a standard yes–no question contour ends in ***H*** – ***H***%. Five types of pitch accent occur in the AmE–ToBI for American English: two simple accents (***H****** and ***L****, and three complex ones, ***L****** + ***H***, ***L*** + ***H****** and ***H*** + !***H****). As in Pierrehumbert's system, the asterisk indicates which tone is aligned with the stressed syllable of the word bearing a complex accent. Unlike in Pierrehumbert's system, there is no ***H**** + ***L*** accent. Also, the ***H***– phrase accent and each of the pitch accents with a ***H*** tone in it can additionally be marked as the start of a downstep register; so, !***H**** is the downstepped counterpart to ***H****, ***L*** + !***H**** is the downstepped counterpart of ***L*** + ***H****, and so on.

[1] A fuller description of the ToBI systems may be found in the ToBI conventions document and the training materials available at http://ling.ohio-state.edu/\sim tobi. To circumvent the ambiguity between the original American English system and the framework, we will substitute the term "AmE–ToBI" in the remainder of this paper.

1.2. The cost of prosodic labeling

ToBI labeling is still a slow manual process. Even when done by highly trained and experienced labelers using multiple displays of the acoustic signal that are time-aligned with audio to allow interactive audiovisual examination of the portion of speech being tagged, AmE–ToBI labeling commonly takes from 100–200 times real time. That is, a 10-s utterance would require from 17 to 33 minutes to label. Because of the time, effort and expense involved in prosodic labeling, we were motivated to speed up the manual process, but did not wish to do so in a way that would seriously bias the assignment of ToBI labels.

The approach to prosodic labeling described in this paper was inspired by the technique of speeding up manual segmental (phonetic) labeling by using speech technology to automatically assign and align phones to a speech sample. In the case of the commercially available Entropic Aligner (Wightman and Talkin, 1994), for example, a speech recognizer is constrained to a known word string using forced alignment. Entropic waves+ (Waves+ Manual, 1996) style label files are automatically generated listing time-aligned phones and words. Manual correction of the automatically determined labels is then done by phonetic labelers. Using a recognizer can result in productivity gains for transcribing orthography as well; for orthographic transcription, working from draft text produced by a recognizer, the manual transcription time is reduced by 20–25% (Wightman, 1999).

2. Experiment on semi-automatic prosody labeling

We developed a semi-automated prosody labeling technique designed to speed manual AmE–ToBI labeling. We conducted an experiment to determine whether labeling by manual correction of automatically generated default AmE–ToBI labels was in fact faster, and whether the technique resulted in bias in the assignment of labels. In this section we first describe our semi-automated technique of labeling prosody, and then describe the design and results of the experiment intended to evaluate that technique.

2.1. Automatic assignment of default labels

We developed a technique to semi-automate prosodic labeling; first, default AmE–ToBI labels were automatically assigned; the default labels were then manually corrected by experienced labelers. Instead of using ASR technology, as did Aligner for segmental labeling, we used TTS technology to predict prosody based on text. Because a system for prediction of word accent and phrase boundary location had been developed previously as part of a TTS system, the extension of this system to provide utterance-aligned default labels required only a minimal investment of experimenter time.

2.1.1. Pitch accent and phrase boundary prediction from text

Default pitch accents and phrase boundaries were based on prosodic predictions made (solely) from text by the prosodic assignment modules of the AT&T NextGen TTS system (Beutnagel et al., 1999). These full assignment modules had been trained on several hand-labeled speech corpora, using features of text analysis of the transcription to predict prosodic labels, a process described in (Sproat et al., 1992; Hirschberg, 1993). Previous evaluation of the algorithms implemented in these modules indicated that, for binary decisions regarding presence versus absence of pitch accent and presence versus absence of major prosodic boundary, the algorithms predicted accent location correctly in just over 82% of cases for read news stories, 85% for a spoken dialogue systems corpus, and 98% for short laboratory sentences (Sproat et al., 1992; Hirschberg, 1993). Phrase boundary location was predicted correctly in 95% of cases, when evaluated on a large corpus of news stories (Wang and Hirschberg, 1992; Hirschberg and Prieto, 1996). No evaluation of these modules' assignment of accent or boundary type has been made. Consequently, in the current study, we used only the binary TTS decisions, 'This word is accented or not' and 'There is an intonational phrase boundary between these two words or not', in assigning default prosody. The discussion below is of this simplification of the TTS accent and phrasing assignment process.

Pitch accents were predicted from text using hand-crafted rules derived from analysis of prosodically hand-labeled speech corpora; automatically-derived rules using the same features have performed with approximately the same accuracy on test data. Both types of rules make use of simple distance measures for the word whose accent status is to be predicted: distance of the word from the beginning and end of the sentence, distance from prior and subsequent intonational phrase boundary, and total words in sentence. They also make use of inferred part-of-speech and morphological information to assign input word tokens to one of four broad classes – closed-cliticized, closed-deaccented, closed-accented and open – based upon frequency distributions in the training data. [2] Membership in one of these word lists is used as a feature in the accent assignment process.

Then, for each word to be assigned accent status, the following additional information is collected: preposed adverbials are identified from surface position and part-of-speech, as are fronted prepositional phrases, and labeled as potentially CONTRASTIVE. CUE PHRASES (discourse markers, such as "well" and "now" which provide explicit structural information about the text) are identified from surface position and part-of-speech, and their accent status is predicted following empirical findings on the textual and intonational disambiguation of cue phrases in (Litman and Hirschberg, 1990). [3] Verb-particle constructions are identified by table look-up.

[2] These categories, intuitively, dividing words into closed (function words, e.g. prepositions and articles) and open (content words, e.g. nouns and verbs) classes, represent results of an earlier study of several speech corpora with respect to the probability of accenting, deaccenting or cliticizing (deaccenting, reducing the vowel, and eliminating word boundaries with prior and subsequent words) members of these two broad classes. Corpus statistics were used to group words into one of four categories: closed-cliticized (function words which were cliticized most of the time), closed-deaccented (function words that were deaccented but not cliticized), closed-accented (function words most frequently accented) and open class words (Hirschberg, 1993).

[3] Lexical items which function as cue phrases generally may also have a more 'semantic' function, as when "well" or "now" serve as adverbials.

An approximation of local focus (roughly, the current topic of discussion, cf. (Grosz and Sidner, 1986)) is implemented as a stack of lemmas of all content words in a phrase. A sentence such as 'The children are reading their assignments.' would produce the following lemmas, to be added to the focus stack: [child, read, assign]. New sentences cause additional items to be pushed onto the stack, which functions as a kind of discourse history. We make use of this history to infer whether words are to be treated as 'given' (old in the discourse) or 'new' (Prince, 1992). Given items are treated as potentially deaccentable (Ladd, 1979; Nooteboom and Terken, 1982). Cue phrases trigger either push or pop operations on the stack, to manipulate the discourse history, roughly as described in (Grosz and Sidner, 1986). Paragraph boundaries cause the entire stack to be popped. In effect, this results in variation of which items will be treated as potentially deaccentable, based upon an inference of the discourse structure of the text to be synthesized.

Noun compounds and their citation-form stress assignment are identified, again by building rules by hand, based upon analysis of a large corpus, as described in (Sproat et al., 1992; Liberman and Sproat, 1992). These rules make use of part-of-speech information and predictions of the semantic category of elements of the complex nominal, and are used to propose a default accent pattern for each complex nominal (e.g. distinguishing between the left-stress pattern of "Main Street" and the rightstress pattern of "Park Avenue"), which may be overridden by discourse-level information, as noted below.

Finally, possible contrastiveness is inferred by comparing the presence of lemmas of component parts of a nominal that is present in the focus stack; if some items are 'given' and others 'new', the new items are marked as potentially contrastive. Accent assignment is then determined as follows: assign 'closed-cliticized' and 'closed-deaccented' items the status accorded their class. Next, 'contrastive' items are accented. Then 'closed-accented' items are accented, remaining 'given' items are deaccented, remaining noun-compound elements are assigned their citation-form stress pattern, and all other items are accented. The algorithm implemented is shown in

For each item w_i labeled with part-of-speech p_i:

If w_i is a phrasal verb, deaccent;
Else if p_i is classified 'closed-cliticized', cliticize;
Else if p_i is classified 'closed-deaccented', deaccent;
Else if w_i is marked 'contrastive', 'prefixed', or 'preposed', accent;
Else if w_i is labeled a cue phrase, accent;
Else if w_i is part of a proper nominal
    If w_i's status is 'given', accent,
    else accent;
Else if w_i is in global focus but not in local focus ('given'), accent;
Else if w_i is classified 'closed-accented', accent;
Else if w_i is in local focus ('given'), deaccent;
Else if w_i is part of a (common) complex nominal
    If w_i is predicted to be accented in citation form, accent
    else deaccent;
Else accent w_i.

Fig. 1. Accent assignment algorithm.

Fig. 1. In each case, accented items were assigned a default ***H**** accent.

Previous experiments reported that this algorithm models binary human accent decisions in radio speech with about 82% accuracy, in short read sentences with 98% accuracy, and in spontaneous elicited speech (the DARPA ATIS corpus) with about 85% accuracy (see Hirschberg, 1993).

Intonational phrase boundaries are predicted using these accent predictions, together with part-of-speech information, punctuation, measures of sentence length and distance of potential boundary from the ends of the sentence and from punctuation earlier in the sentence. The full set of features includes, for each potential phrase boundary between two words $\langle w_i, w_{i+1}\rangle$:

- a part-of-speech window of four around the site, $\langle w_{i-1}, w_i, w_{i+1}, w_{j+2}\rangle$;
- whether w_i and w_{i+1} bear a pitch accent or not;
- the total number of words in the sentence;
- the distance in words from the beginning and end of the sentence to $\langle w_i, w_{i+1}\rangle$;
- the distance in syllables and in stressed syllables of $\langle w_i; w_{i+1}\rangle$ from the beginning of the sentence;
- the total number of syllables in the sentence;
- whether the last syllable in w_i is phonologically strong or weak, based on lexical stress assignment;
- the distance in words from the previous internal punctuation to w_i;
- the identity of any punctuation occurring at the boundary site;
- whether $\langle w_i, w_{i+1}\rangle$ occurs within or adjacent to an NP;
- if $\langle w_i, w_{i+1}\rangle$ occurs within an NP, the size of that NP in words, and the distance of $\langle w_i, w_{i+1}\rangle$ from the start of the NP.

A classification and regression tree (CART) (Breiman et al., 1984) analysis was performed on a corpus of approximately 89,000 words of AP news text, hand-labeled by a native speaker for likely full intonational boundaries (Hirschberg and Prieto, 1996). This analysis was used to construct decision trees automatically from the variables described above. The CART cross-validated estimate of the generalizability of the tree grown from this data was 95.4%; that is, the CART method predicted this success rate for unseen data, given this tree. [4] The tree was used to generate a binary

[4] CART cross-validation 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 subtree 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 the subtrees.

phrase-break decision (full intonational boundary versus no boundary), for the utterances labeled in the current study, from the input transcription. Sentence-internal boundaries were labeled ***L – H%*** (the most common internal full intonational phrase boundary identified in previous corpora) and (transcription) sentence final boundaries were labeled ***L – L%***, unless the sentence was identified as a yes–no question; this tagging is done by a simple procedure which uses punctuation and initial key-words to spot questions and distinguish wh-questions from yes–no questions. No intermediate boundaries are predicted by this procedure, so none were generated in the prosodic hypotheses.

Together, the accent and phrasing modules predict which words are to be accented and where intonational phrase boundaries are to be placed in the TTS system. These procedures currently assign accents and phrasing in both the Lucent Bell Labs TTS system and the AT&T Labs NextGen TTS system. When applied to the orthographic transcription of the read speech in our corpus, these prosodic predictions formed an initial labeling hypothesis for our AmE–ToBI labelers.

2.1.2. Automatic default AmE–ToBI label assignment and alignment

Simple heuristics were developed and implemented in a Perl (Wall et al., 1996) program to determine the AmE–ToBI label assignment and alignment to the speech signal. Inputs to the program included:

- A file (in Entropic waves+ format) listing the time-aligned words spoken in a speech file (i.e., the orthographic tier).
- The output from the accent and phrasing modules described above, which listed each word in the utterance and its automatic classification with respect to the presence of absence of accent, cliticization, and following phrase boundary.
- A file containing the punctuated text of the speech file after being normalized by the TTS text normalization module. The text normalization module expands abbreviations and non-alpha character input such as digits and symbols into English words, and also identifies and flags yes–no type questions.

The heuristics for automatic assignment of default breaks, pitch accents and phrase accents and boundary tones may be summarized as follows:

Breaks (aligned to end of word):
  If clitic, break = 0;
  Else if phrase boundary, break = 4;
  Else break = 1.
Pitch Accents (aligned to middle of word):
  If accented, pitch accent = ***H****.
Phrase Accents and Boundary Tones (aligned to end of word):
  If phrase boundary and
    If sentence internal, edge tone = ***L – H%***;
    If punctuation is flagged as yes–no question, edge tone = ***H – H%***;
    Else, edge tone = ***L – L%***.

Note that since the default labels were generated solely from text without the use of acoustic information, alignment could not correspond to f_0 maxima or minima within the syllable predicted to be accented. Labelers were expected to move the default labels to the appropriate location. Had alignment been made instead to the middle of the primary stressed syllable, it is possible that labeling time could have been further decreased.

2.2. Speech corpora

All the speech labeled during the experiment was read by one professional female speaker. It composed some of the acoustic inventory for an AT&T concatenative synthesis TTS system. Two different prosodic styles were compared in the experiment: business news reading (to be identified as the WSJ corpus because of the many Wall St. Journal articles it contains), composed of sets WSJ1 and WSJ2, and interactive service prompts from various AT&T telephone service applications (to be termed the Prompts corpus, composed of sets P1, P2 and P3), which contained greetings, questions, instructions, apologies, etc. The WSJ corpus consisted of 61 minutes of speech and 7,677 words, and the Prompts corpus, 29 minutes and 4,448 words. Another corpus (to be termed set L1 (Lab1)), composed of short laboratory sentences chosen for their phonetic coverage, was also transcribed; this corpus consisted of 30 minutes of

speech and 5,318 words. The purposes of including different corpora were to explore the prosodic differences between the two styles and also to see whether prosodic style affected the labeling process.

2.3. Labeling procedures

2.3.1. ToBI transcription scripts

Transcriber scripts were written for the labelers to use in labeling the speech files in two different labeling modes – "Scratch" and "Default" (see below). These scripts also automatically logged the times when a file was opened and closed in any given labeling session. The scripts displayed the speech files using Entropic xwaves+. The sound file and f_0 file were time aligned with six tiers of labels consisting of the four standard tiers and two site-specific tiers, for (1) tones, (2) breaks, (3) words, (4) syllable boundaries (to aid in placing accent marks correctly on the tones tier), (5) miscellaneous events such as dysfluencies and (6) comments where the labeler could flag points of potential ambiguity to be discussed with the other labelers in group meetings. The display is illustrated in Fig. 2, which shows what the screen looked like when labeling in the Default labeling mode. When labeling from Scratch, the words file for each utterance was first copied into the breaks files so that breaks and words would be automatically aligned, and so that the break index labels could be inserted by simply replacing each word label in the breaks tier with a break index label (0–4) accessible via a mouse-driven menu. Note that utterance-aligned orthographic words files were used by the Scratch method as well as the Default method, so the only possible advantage for the Default method was tone prediction and placement and break index prediction.

An awk (Aho et al., 1988) "checker" script written by John Pitrelli for an earlier intertranscriber agreement study (Pitrelli et al., 1994) was used to check the transcription of tones and breaks. When run over the label files produced by the transcriber scripts, this program ensures that the tonal analysis is complete and "grammatical". For example, for each phrase accent, there must be at least one preceding pitch accent label within the intermediate phrase (i.e., after the immediately preceding phrase accent). If there is no such pitch accent label, the error and its location relative to the beginning of the sound file is returned, so that changes can be made. The program also checks that there is agreement between labels in the tones tier and the breaks tier. Each label of 4 in the breaks tier, for example, must be accompanied by a full intonational boundary tone in the tones tier. After doing a transcription, the labeler ran the "checker" program, and then re-labeled to correct any flagged errors, before running "checker" again, in an iterative process until no further errors were returned. Correction was done using the same scripts, so that the time that it took to correct checker-flagged errors also could be logged as part of the labeling time.

2.3.2. Assignment of labeling

Sets of utterances (including f_0 files and sound files with aligned word and syllable labels) were provided by AT&T Labs on a weekly or bi-weekly basis to the labelers. These files were distributed among the labelers so that all of the labelers working on the project at any stage could have an approximately equal work load. More critically for the current study, the set of utterances assigned to each labeler was divided into two parts with equal amounts of speech in each part. One half was designated to be transcribed from "Default" and the other from "Scratch". The script for labeling in Default mode brought up label files that included a complete tonal transcription and break indices that had been automatically generated as described in Section 2.1. Thus, when labeling in Default mode, the labeler could often simply replace any labels with which he or she disagreed using the application menu. The script for labeling in Scratch mode brought up no tones or breaks labels (except that the location of the breaks was provided – see the description of the labeling scripts above). Each labeler thus had to assign all of the tone and breaks labels for the file. Labelers were instructed to complete the default and scratch files in alternating order, and care was taken in the assignment to ensure that equal amounts of speech data were analyzed in each of the two labeling

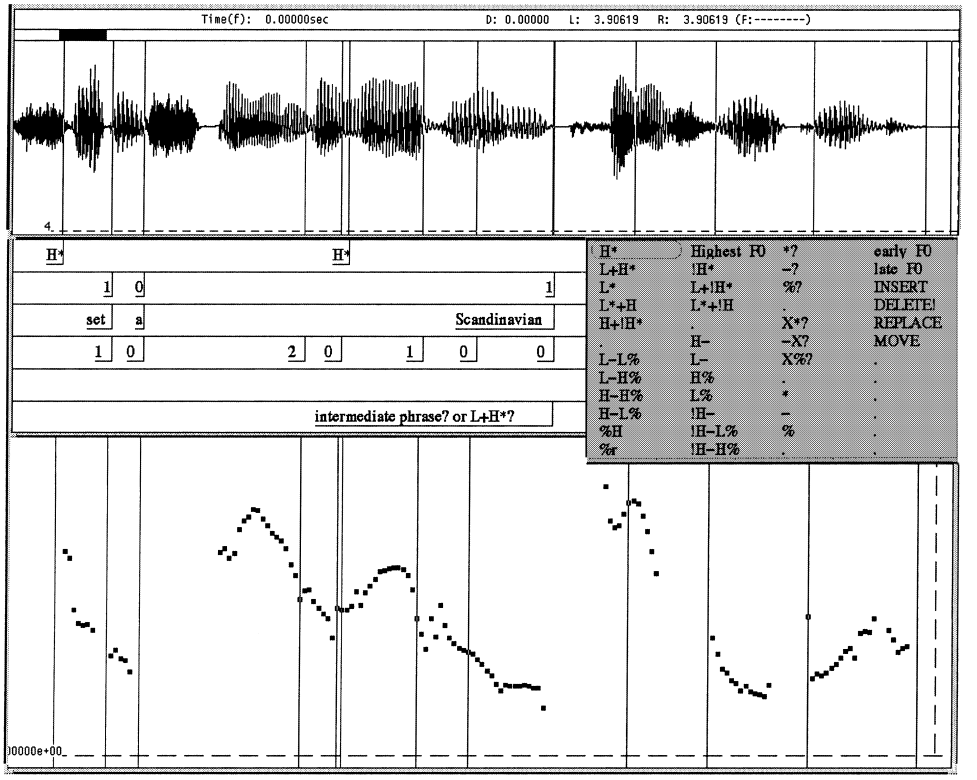

Fig. 2. Example of display screen for AmE–ToBI labeling.

modes at any stage of labeler experience (see next section).

2.3.3. Labeler experience and training

Five labelers – JM (the third author) and LM, KB, CH and AM (four doctoral students in Linguistics at Ohio State University) – were employed by AT&T Labs – Research to transcribe the intonation patterns of utterances within the current investigation, which lasted nine months. Transcriptions included a complete analysis on intonation patterns and break indices using the AmE–ToBI system, as described in the "ToBI Annotation Conventions" (Beckman and Hirschberg, 1994) and the accompanying "Guidelines for ToBI Labeling" (Beckman and Elam, 1997). All labelers had worked through the "Guidelines to ToBI Labeling" before starting to transcribe, and both JM and KB had some prior labeling experience. Two labelers, JM and LM, labeled during the entire study; KB worked on the initial five corpus sets; CH and AM replaced KB for the last corpus set (WSJ2) labeled during the study. All of the labelers had access to the expertise of one the developers of the "Guidelines ..." (MB, the fourth author of this paper), who collaborated with JM in supervising the project.

JM had been involved in AmE–ToBI labeling of other portions of the AT&T corpus for six months when the other labelers began transcribing utterances. Because of her more extensive experience, JM was assigned to train each of the four other labelers, and she oversaw all work to ensure greater consistency between labelers. Training took place iteratively during the first two weeks of a labeler's assignment to the project. After being initiated into the use of the two labeling scripts (see above), a labeler labeled several files with JM present to answer questions and point out errors in the transcription. After several sessions of such one-on-one tutoring, the labeler then worked alone on several files, before the next tutorial

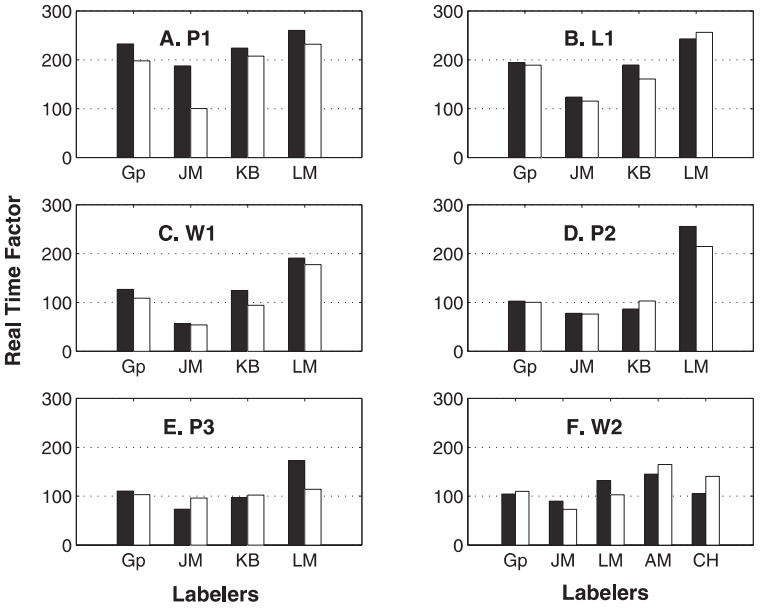

Fig. 3. Scratch (dark bars) versus default (light bars) labeling rates per corpus subset for group (Gp) and individual labelers.

session, during which JM again targeted errors and answered questions regarding these files. Particular emphasis was placed on imparting criteria for deciding between label pairs that have been the locus of relatively lower agreement levels in previous inter-transcriber consistency tests such as (Beckman and Elam, 1997). These included pitch accent pairs $L^* + H$ versus $L + H^*$ H^* versus $L + H^*$ and H^* versus $!H^*$. The edge tone combinations of $L - L\%$ versus a single $L-$ and $H - H\%$ versus $H-$ were also potentially confusing and so were also reviewed. In addition, edge tone combinations that are covered less thoroughly in the "Guidelines ..." were reviewed. These included $!H - L\%$ and $!H - H\%$. Each labeler spent 20 hours in this initial stage. The files labeled during training are included in the analyses of the two labeling modes reported below. Thus the labeling times in panel A in Fig. 3 reflect the training times for labelers KB and LM, and the labeling times in panel F reflect training times for labelers AM and CH. In training KB and LM, extreme care was taken to ensure that the training sets included an equal amount of labeling in each of the Scratch and Default labeling modes. Experimenter error resulted in less good control of the alternation between Scratch and Default in training AM and CH (see Table 1). Nevertheless the amount of speech labeled using each method was equivalent over all labelers (Default: 944 s; Scratch: 941 s) for subset W2.

After training, labelers continued to meet as a group every week or two, to discuss any questions or problem cases in the batch of files just transcribed. MB also participated in these group meetings, to help JM resolve any particularly difficult ambiguities. In preparation for these group meetings, the labelers recorded utterance-specific questions in the comments tier, and more general questions (along with file names for specific examples) in a group journal. Thus, the group meetings in effect helped to calibrate uniform labeling conventions for unusual style-specific patterns, such as $L + H^*$, $!H^*$, H^* for a pattern common in the Prompts.

3. Results

Our analysis of results focused on two questions:

1. Was the method of manual correction of automatically assigned Default labels faster than

standard manual prosody labeling "from scratch"?
2. Was any bias in label assignment introduced by the Default method?

3.1. Labeling time

The labeling time of a speech file was normalized by representing it as a multiple of the duration of the speech file. The labeling time multiple will be referred to as the real time factor (RTF = labelingtime ÷ speechtime). Using RTFs for each of the 962 speech files labeled as input data, a standard two-sample t-test of the null hypothesis that the Scratch RTF was greater than the Default RTF was marginally significant ($t = 1.6314$, df $= 960$, $p = 0.0516$). When subset W2 was omitted from the analysis, the difference was significant ($t = 1.657, \mathrm{df} = 915$, $p < 0.05$). Table 1 lists, for each labeling method, the total labeling time, total duration of speech files, and overall RTF for each of the individual labelers and for the group. Each of the three major labelers (JM, KB and LM) reflect the group results, namely that labeling from Default was faster than labeling from Scratch. For the two less experienced labelers (AM and CH, who labeled only for the final subset of the corpus), however, the Scratch labeling method was faster than the Default method.

Fig. 3 is composed of six bar plots of real time factors for subsets of speech files labeled in Scratch and Default modes; results from each corpus subset are represented by a separate plot. The bar plots are ordered from the first subset labeled (Prompts 1 (P1)), shown in panel A, to the last (WSJ2), in panel F. The results are shown for the group ("Gp") and for each labeler individually. The effect on labeling rate of experience with the speaker and with the task is evident in the reduced real time factors for both Scratch and Default cases as the study progressed. Each of the three major labelers (JM, KB and LM) gradually increased their labeling speeds; their individual RTFs were reduced eventually to about half their initial values. It is also apparent from the Group data that labeling from Defaults was faster than labeling from Scratch for five of the six speech subsets. For the major labelers, labeling from Default was faster than from Scratch across most, but not all, speech subsets. Major labelers JM and LM reversed the overall effect for one of the subsets, and KB reversed the effect for two, although which subsets were reversed varied among the labelers.

Since there were relatively greater savings in labeling time with the Default method for subsets P1 and W1, in which the speech files contained lengthy paragraphs, than for subsets containing shorter utterances, we initially hypothesized that the longer the speech file to be labeled, the more the Default method speeded labeling. To test this hypothesis, RTF was plotted as a function of speech file duration for Scratch and Default cases (see Fig. 4). The very similar scatter plots, how-

Table 1
Labeling time, speech duration and real time factor (RTF) for speech files labeled by each labeler and the group using each method

Labeler	Method	Labeling time (s)	Speech time (s)	RTF
JM	Scratch	98,184	1046.91	93.78
JM	Default	87,346	1068.16	81.77
KB	Scratch	166,810	1052.73	158.46
KB	Default	137,004	996.80	137.44
LM	Scratch	196,941	913.87	215.50
LM	Default	182,647	872.33	209.38
AM	Scratch	6,625	45.69	144.99
AM	Default	21,904	132.71	165.05
CH	Scratch	41,416	393.30	105.30
CH	Default	41,641	295.85	140.75
Group	Scratch	509,976	3452.51	147.71
Group	Default	470,542	3365.85	139.80

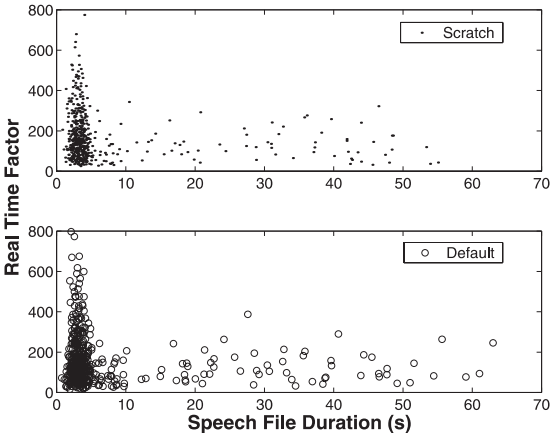

Fig. 4. Scratch (top) and default (bottom) RTF labeling rates as a function of speech file duration (in seconds).

ever, indicate that RTF does not increase with speech file duration for either case. Correlations, in fact, indicated a slight tendency in the opposite direction (−0.13 for Scratch and −0.11 for Default).

Looking closer at the results in Fig. 3, we note that JM and KB quickly became considerably faster labelers as their experience increased, and for them, the biggest differences between the two modes was for subsets labeled relatively early in the study. The apparent reduction or reversal of the overall effect for them may be a kind of ceiling effect. LM's patterns, on the other hand, suggest that she was a more cautious labeler, whose labeling speeds benefited from the Default mode more as she became used to each new speaking style. However, the current study does not include enough different subsets of enough different styles to explore this suggestion of a "speed-accuracy tradeoff" effect in individual labeling styles.

3.2. Prediction accuracy of defaults

In this section, we present a brief analysis of the accuracy of the predicted prosodic labels for the speech corpora used in the current study. The accuracy of the predicted default tones and breaks would be expected to have a fairly straightforward influence on the speed of labeling using the Default method. The more accurate the default labels, the faster the manual labeling process, since there would be fewer corrections to be made. It is less clear how default accuracy might also affect labeling bias. On one hand, highly accurate labels might introduce bias because labelers would have more confidence that they are correct, and might pay less attention to searching for errors. On the other hand, more accurate defaults could give labelers relatively more time to identify the fewer errors, and thus decrease labeling bias.

Table 2 lists the word level accuracy of default labels in making binary predictions about whether or not a word receives a pitch accent, and whether or not a word marks an intonation phrase boundary. Only binary predictions were considered, because it was not our intent in the current study to predict types of pitch accent or of phrase boundary. Rather, we wanted as an initial step to see whether relatively simple but robust predictions of prosodic phenomena based on already existing text analysis algorithms might assist in manual prosodic labeling. Overall accuracy is defined as the percentage corresponding to the sum of the number of correct predictions (e.g., correct predictions of both accent and of no accent) divided by the total number of words.

Overall prediction accuracy of both accent and phrase boundary location was lowest for the Prompts corpus. Unlike the current study,

Table 2
Accuracy of default accent and phrase location prediction for various speech corpora

Corpus	Pitch accent: % correct			Major phrase: % correct		
	Overall	Accent	No accent	Overall	Phrase	No phrase
Lab	88	91	84	87	78	88
Prompts	79	74	85	83	80	84
WSJ	89	95	82	90	79	92
All	86	89	83	87	79	89

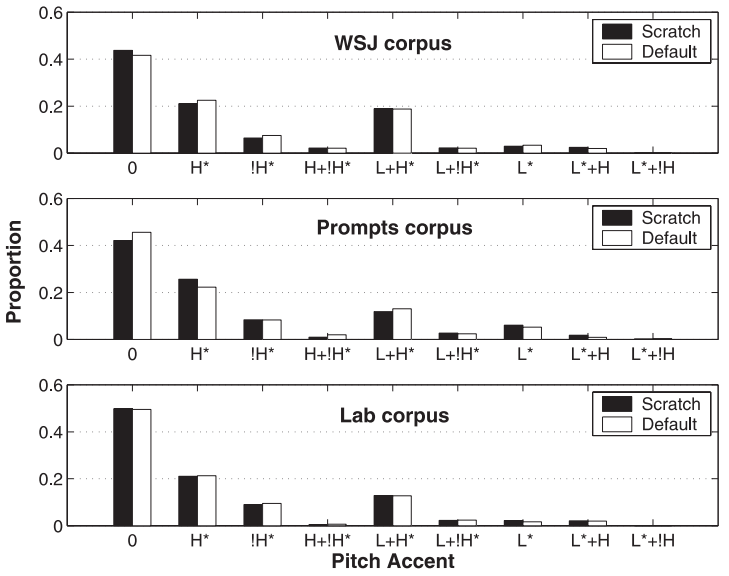

Fig. 5. Pitch accent distributions by labeling method and corpus.

previous evaluations of pitch accent placement (Sproat et al., 1992; Hirschberg, 1993) found slightly higher accuracy for a spoken dialogue system corpus (like the Prompts corpus) than for read news stories (like the WSJ corpus), and the highest accuracy for short laboratory sentences (like the Lab corpus). In the current study's Prompts corpus, there were a disproportionate number of errors (26%) of accented words that were predicted to be unaccented; this is a "miss" type of error. Errors of this type in the other corpora were under 9% of accented words. There was less difference among corpora in phrase boundary prediction accuracy. The major type of phrase boundary error was that a large percentage of words that actually marked a phrase boundary (21% in the entire corpus) were not predicted to do so; this is also a "miss" type of error.

3.3. Distribution of labels

To test for the possibility of bias in labeling introduced by the Default method, the distributions of AmE–ToBI labels for Default and Scratch methods were compared. Separate comparisons were made for the three corpora representing different prosodic styles: Lab, Prompts and WSJ. Distributions of pitch accents, phrase tones and break indices were each compared.

3.3.1. Pitch accents

Fig. 5 illustrates the proportion of words in the WSJ, Prompts and Lab corpora that were manually assigned each of eight pitch accent types plus the assignment of no accent for Default and Scratch labeling methods. Since all words predicted to receive pitch accents were assigned ***H**** by default, any bias effects should at least affect unaccented words or ***H**** accents, although we tested the entire distribution. For the WSJ corpus, no significant difference was found between the distributions of pitch accents for Default and Scratch methods (Pearson's chi-square test, $\chi^2 = 10.6964,\ \mathrm{df} = 7,\ p = 0.1524$). [5] We interpret this result to indicate that no bias in pitch accent assignment was introduced by the Default method

[5] For the chi-square test, accents ***L**** + ***H*** and ***L**** + !***H*** were combined into one category, since expected counts were <5 for the latter accent, which could produce inappropriate χ^2 approximations. Categories were collapsed in some subsequent chi-square tests as well, or in rare cases they were omitted from the analysis, if cell counts were <5 but the cell could not be collapsed reasonably with another category.

for the WSJ corpus. When the chi-square test was applied to the subset W1 alone, the results were the same ($\chi^2 = 8.3485$, df $= 7$, $p = 0.3029$). In addition, none of the chi-square tests conducted for each of the three major labelers individually found any significant differences between Scratch and Default pitch accent distributions in the WSJ corpus. Thus, the apparent lack of bias here cannot be attributed to the less good control of the alternation between the two labeling modes with the two new labelers who replaced KB for corpus W2.

For the Prompts corpus, the distributions of pitch accents were significantly different between the Default and Scratch methods ($\chi^2 = 26.539$, df $= 8$, $p < 0.001$). This statistic means that if the Scratch and Default samples were actually from the same larger distribution, the probability of selecting, by chance alone, two samples that differed as much as the observed samples did is less than one in 1,000. The result can be attributed to the difference in the proportions of words given no pitch accents and those assigned ***H**** accents for the two labeling methods. As can be seen in Fig. 5, a higher proportion of words were unaccented for the Default labeling method (46%) than for the Scratch method (42%). Across all corpora, the percentage of words assigned no pitch accent by default (46%) was slightly higher than the percentage of words manually assigned none (45%) using the Scratch method. Labelers' accent placement decision appears to have been biased by the Default method, but only for the Prompts corpus. We speculate that the disproportionately high number of misses in predicting accented words (that was characteristic of the Prompts corpus) influenced labelers to accent fewer words when they were labeling by the Default method than they otherwise did when labeling from Scratch. Although each of the three primary labelers assigned proportionally more words to the unaccented (0) category when labeling the Prompt corpus from Default than from Scratch, there were no significant differences between their Scratch and Default pitch accent distributions when tested individually.

For the Lab corpus, there were no significant differences between Scratch and Default pitch accent distributions either for the entire corpus ($\chi^2 = 2.7383$, df $= 7$, $p = 0.9081$), or for each of the three labelers individually.

3.3.2. Edge tones

No significant differences were found between the distributions of edge tones for Default and Scratch methods for the WSJ, the Prompts or the Lab corpus (Pearson's chi-square test, $\chi^2 = 9.9172$, df $= 8$, $p = 0.2709$ for WSJ; $\chi^2 = 9.1547$, df $= 9$, $p = 0.4231$ for Prompts; $\chi^2 = 3.5673$, df $= 6$, $p = 0.735$ for Lab). Thus, no bias in labeling edge tones appears to have been introduced by the Default method.

3.3.3. Break indices

Significant differences in the distributions of breaks were observed between the Default and Scratch labeling methods for the full WSJ ($\chi^2 = 40.7317$, df $= 4$, $p < 0.0001$), Prompts ($\chi^2 = 113.1865$, df $= 4$, $p < 0.0001$) and Lab corpora ($\chi^2 = 64.7701$, df $= 4$, $p < 0.0001$). However when the W1 subset of the WSJ corpus was analyzed independently, there was no significant difference between Scratch and Default break distributions ($\chi^2 = 6.3487$, df $= 4$, $p = 0.1746$), even though the W1 subset was comparable in size to the Prompts and Lab corpora, and hence the tests were similar in statistical power. As can be seen in Fig. 6, there was a higher percentage of 0 breaks for the Default method than for the Scratch method for all corpora. For WSJ, 10.5% of the breaks were labeled 0 using the Scratch method, as compared to 14.8% for the Default method; for Prompts, the percentages were 5.7% for the Scratch method and 15.7% for the Defaults method; for the Lab set, 7.0% for Scratch and 13.6% for Defaults. Since fully 18% of words were assigned a default break index 0, it appears that the Default labeling method biased labelers towards labeling substantially more 0 breaks than they otherwise would do when labeling from Scratch.

Scratch and Default break distributions were also compared for each of the three primary labelers individually. For the Prompts corpus, each labelers' Scratch versus Default break index distributions differed significantly and followed the same pattern as the Group results (JM: $\chi^2 =$

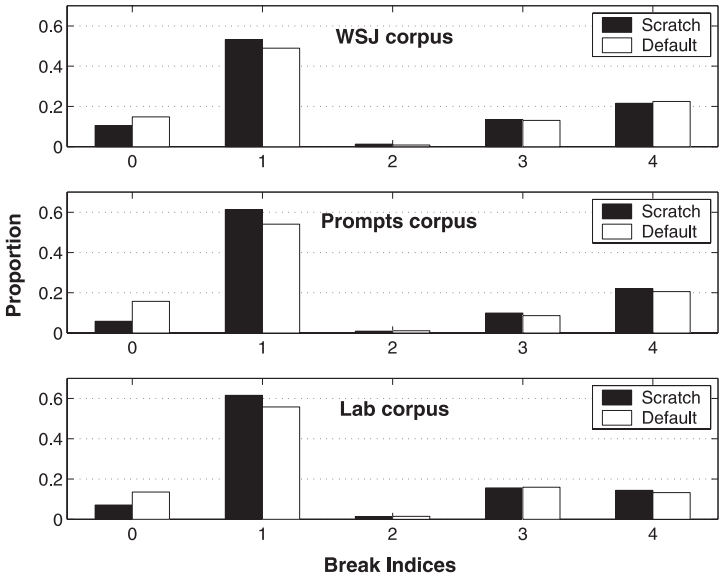

Fig. 6. Break distributions by labeling method and corpus.

38.6823, df = 4, $p < 0.0001$; KB: $\chi^2 = 69.1442$, df = 4, $p < 0.0001$; LM: $\chi^2 = 19.6552$, df = 4, $p < 0.001$). Similarly for each of the three labelers of the Lab corpus, Scratch and Default break index distributions differed significantly (JM: $\chi^2 = 10.8541$, df = 3, $p < 0.05$; KB: $\chi^2 = 33.3358$, df = 4, $p < 0.0001$; LM: $\chi^2 = 31.9378$, df = 4, $p < 0.0001$). For the WSJ corpus, however, only labeler JM's distributions of breaks were significantly different between Scratch and Default labeling methods ($\chi^2 = 37.5135$, df = 4, $p < 0.0001$). This result is interesting in view of the fact that JM was both the most experienced and the fastest ToBI labeler of the group. This suggests that labeling speed rather than lack of experience may be the better predictor of which labelers are more prone to bias from Default labels for break indices.

It appears that labeling from Default typically introduced bias in favor of break index 0, although less so for the WSJ corpus than for the Lab or Prompts corpora. To put this apparent bias into perspective, however, the reliability of labelers needs to be considered. In another study conducted with the same group of labelers, inter-transcriber agreement was only 42.3% for break index 0 when each labeler in the group was labeling a large common set of utterances from Scratch (McGory et al., 1999). When one labeler transcribed a break index of 0 for a word, another labeler from the group was more likely to assign it a break index of 1 than to also label it 0. Inter-transcriber agreement for 0 breaks with the Defaults method (51.6%) was considerably higher than for the Scratch method, but still relatively low compared to agreement for more reliably assigned breaks 1 (Scratch: 83.6%, Defaults: 81.1%) and 4 (Scratch: 94.0%, Defaults: 91.1%). Since the distinction between breaks 0 and 1 was not reliable across labelers using either method, the introduction of bias in favor of break index 0 by the Default labeling method does not seem to be a very serious drawback.

4. Summary and conclusions

ToBI labeling from default labels that were automatically generated from text analysis was found to speed prosodic labeling. The proportion of labeling time saved did not depend on the style of text or on labeling experience in any straightforward way. A comparison across the three main labelers suggests that time saved may depend on

the individual labeler's level of experience interacting with the labeler's personal labeling style. However, there were not enough labelers or subsets of data in different styles to pursue this suggestion further in the current study.

There were two indications of relatively minor biases in AmE–ToBI labeling introduced by the semi-automated Default labeling method. The first involved a higher proportion of unaccented words and a lower proportion of ***H**** pitch accents in the Prompts corpus for the Default method than for the Scratch method. No such bias was evident in the WSJ or Lab corpora, however. In the Prompts corpus, the proportion of words predicted to be unaccented that were in fact accented was three to five times higher than for the other corpora. The unusually large number of missed accent errors in the Default labels of the Prompts corpus appears to have biased labelers to assign fewer accents to words than they otherwise did. This implies that listener bias can be induced by either low accuracy in Default predictions, or by an accuracy rate that is strikingly different for one corpus than for the others also being labeled.

The second instance of bias introduced by the Default labeling method was an increase in the proportion of 0 break indices in all corpora when labeling from Defaults, although the effect was smallest for the WSJ corpus. Since labelers did not label 0 breaks very reliably, however, this bias was not considered to be a serious problem. A possible explanation for the bias and lack of reliability in labeling break index 0 is that linguistic research on levels of juncture in American English has concentrated more on the higher, tonally-marked levels than on segmental markings. For example, it has still not been established what the distribution of the "flap" allophone of /*t*/ is relative to prosodic organization. One possible explanation of the 0 break index result is that the "ToBI Annotation Conventions" and the "Guidelines ..." are much clearer (because they are more solidly grounded in research) on levels 2, 3 and 4 than on levels 0 and 1. This implies that less well-defined prosodic categories are likely to be unreliably labeled and to be affected by Default-induced bias.

4.1. Future directions

It seems reasonable to suppose that adding acoustic information to text-based methods of prosody prediction would result in improved prediction/recognition accuracy. Better accuracy would likely translate to faster semi-automated labeling, and possibly to less biased labeling as well, although the present study did not test a wide enough range of Default accuracy to answer that question. Once an automated system was capable of accuracy equivalent to or better than the inter-transcriber reliability of human labelers, however, it could replace the need for manual prosodic labeling. Because it could rapidly and reliably label large speech corpora, an accurate automated prosody prediction/recognition system would be a boon both to linguistic researchers studying prosody and to speech technology researchers developing speech synthesis and speech recognition systems, both of whom require large accurately labeled speech corpora for their work.

A recent study of prosodic recognition from speech utterances (Conkie et al., 1999) provides some interesting data about the current capabilities of prosody recognition based on acoustic and linguistic models of prosodic events. The recognition study used speech and text data (for training and testing) from the same larger corpus as was labeled in the present study, and focused on the recognition of pitch accent (presence or absence) on a word. The accuracy of the text-based predictions described in the current study provided a baseline with which to compare utterance recognition-based accuracy. Average accuracy (the mean of the accuracies of predicting accent and no accent) for the text-based baseline system was 84.8%, whereas recognition accuracy trained and tested on acoustic information alone (f_0 and energy) was 82.8%. In addition, a part-of-speech (POS) tagger was used on the corresponding text corpus to obtain POS tags for training and testing prosody prediction. Average accuracy for the syntactic system was 84.0%. A combined acoustic/syntactic prosody recognition system, however, achieved average accuracy of 88.3%. The use of text-based predictors other than POS and of additional acoustic information would be expected to

improve accuracy even more. Average inter-transcriber agreement (on a subset of the same corpus) for presence or absence of pitch accents was 91.8% (McGory et al., 1999). Thus automatic techniques combining text-based prediction and acoustic-based recognition of the presence or absence of pitch accents are on the brink of achieving accuracy comparable to human reliability. Future work on automatic techniques is needed, however, to accurately predict/recognize not only the placement but the various types of accents.

Acknowledgements

The authors would like to express their appreciation to the Ohio State University, Department of Linguistics, Graduate student ToBI labelers whose work was described in this paper, and to Volker Strom for his contribution to assessing the prediction accuracy of Default labels.

References

Aho, A.V., Kernighan, B.W., Weinberger, P.J., 1988. The Awk Programming Language. Addison-Wesley, Reading, MA.

Beckman, M.E., Elam, G.A., 1997. Guidelines for ToBI labeling. Guidelines version 3.0, The Ohio State University Research Foundation.

Beckman, M.E., Hirschberg, J., 1994. The ToBI annotation conventions. Appendix A. The Ohio State University Research Foundation.

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., Syrdal, A., 1999. The AT&T Next-Gen TTS System. In: Proceedings of the Joint Meeting of ASA, EAA, and DEGA, Paper 2aSCa4, Berlin, March 1999. J. Acoust. Soc. Amer. 105 (2), 1030 (A).

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks, Pacific Grove, CA.

Byrd, D., 1992. Sex, dialects, reduction. In: Ohala, J.J., Nearey, T.M., Derwing, B.L., Hodge, M.M., Wiebe, G.E. (Eds.), Proceedings of the International Conference on Spoken Language Processing, Banff, October 1992, ICSLP, pp. 827–830.

Conkie, A., Riccardi, G., Rose, R.C., 1999. Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events. In: Proceedings of the European Conference on Speech Communication and Technology. Budapest, September 1999, Eurospeech, ESCA. Vol. 1, pp. 523–526.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S. and Dahlgren, N.L., 1986. The DARPA TIMIT acoustic–phonetic continuous speech corpus CDROM. CDROM, NIST [www.ldc.upenn.edu/lol/docs/TIMIT.html].

Grosz, B.J., Sidner, C.L., 1986. Attention, intentions, and the structure of discourse. Computational Linguistics 12 (3), 175–204.

Hirschberg, J., 1993. Pitch accent in context: predicting intonational prominence from text. Artificial Intelligence 63, 305–340.

Hirschberg, J., Prieto, P., 1996. Training intonational phrasing rules automatically for English and Spanish text-to-speech. Speech Communication 18, 281–290.

House, A., Fairbanks, G., 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. J. Acoust. Soc. Amer. 25, 105–113.

Kenyon, J.S., Knott, T.A., 1953. A Pronouncing Dictionary of American English. G. & C. Merriam, Springfield, MA.

Ladd, D.R., 1979. Light and shadow: a study of the syntax and semantics of sentence accents in English. In: Waugh, L., van Coetsem, F. (Eds.), Contributions to Grammatical Studies: Semantics and Syntax. University Park Press, Baltimore, pp. 93–131.

Lehiste, I., 1960. An acoustic phonetic study of internal open juncture. Phonetica 5, 1–54.

Liberman, M., Sproat, R., 1992. The stress and structure of modified noun phrases in English. In: Sag, I. (Ed.), Lexical Matters. University of Chicago Press, Chicago.

Litman, D., Hirschberg, J., 1990. Disambiguating cue phrases in text and speech. In: Proceedings of the 13th International Conference on Computational Linguistics. Helsinki, August 1990, COLING, pp. 251–256.

Marcus, M.P., Santorini, B., Marcinkiewicz, M.A., 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19 (2), 313–330 [www.cis.upenn.edu/treebank/home.html].

McGory, J., Herman, R., Syrdal, A., 1999. Using tone similarity judgements in tests of intertranscriber reliability. J. Acoust. Soc. Amer. 106, 2242 (A).

Nooteboom, S.G., Terken, J., 1982. What makes speakers omit pitch accents?: An experiment. Phonetica 39, 317–336.

Olive, J.P., Greenwood, A., Coleman, J., 1993. Acoustics of American English Speech: A Dynamic Approach. Springer, New York.

Peterson, G.E., Lehiste, I., 1960. Duration of syllable nuclei in English. J. Acoust. Soc. Amer. 32, 693–703.

Pierrehumbert, J., Hirschberg, J., 1990. The meaning of intonation contours in the interpretation of discourse. In: Cohen, P., Morgan, J., Pollack, M. (Eds.), Plans and Intentions in Communications. MIT Press, Cambridge, pp. 271–312.

Pitrelli, J., Beckman, M., Hirschberg, J., 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In: Proceedings of the Third International Conference on Spoken Language Processing. Yokohama, 1994, ICSLP. Vol. 2, pp. 123–126.

Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S., Fong, C., 1991. The use of prosody in syntactic disambiguation. J. Acoust. Soc. Amer. 90, 2956–2970.

Prince, E.F., 1992. The ZPG letter: subjects, definiteness, and information-status. In: Thompson, S., Mann, W. (Eds.), Discourse Description: Diverse Analyses of a Fund Raising Text. John Benjamins B.V., Philadelphia, pp. 295–325.

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P., Hirschberg, J., 1992. ToBI: A standard scheme for labeling prosody. In: Proceedings of the Second International Conference on Spoken Language Processing. Banff, October 1992. ICSLP, pp. 867–879.

Sproat, R., Hirschberg, J., Yarowsky, D., 1992. A corpus-based synthesizer. In: Proceedings of the International Conference on Spoken Language Processing. Banff, October 1992, ICSLP, pp. 563–566.

Wall, L., Christiansen, T., Schwartz, R., 1996. Programming Perl. O'Reilly, Sebastopol, CA.

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

Waves+ Manual, 1996. Entropic Research Laboratory, Inc.

Wightman, C., 1999. Personal communication.

Wightman, C., Talkin, D., 1994. The aligner: A system for automatic time alignment of English text and speech. Document version 1.7, Entropic Research Laboratory, Inc.