Speech Recognition Models of the Interdependence Among Syntax, Prosody, and Segmental Acoustics

Mark Hasegawa-Johnson, Jennifer Cole, Chilin Shih, Ken Chen, Aaron Cohen, Sandra Chavarria, Heejin Kim, Taejin Yoon, Sarah Borys, and Jeung-Yoon Choi
University of Illinois at Urbana-Champaign
{jhasegaw,jscole,cls,kenchen,ascohen}@uiuc.edu
{chavarri,hkim17,tyoon,sborys,choijy}@uiuc.edu

Abstract

This paper describes results from several dozen experimental systems, and draws conclusions about the ability of speech recognition models to represent the relationship among syntax, prosody, and segmental acoustics. Prosody-dependent allophone modeling can reduce the word error rate (WER) of a speech recognizer, but only if both the language model and the acoustic model encode explicit dependence on prosody. Word error rate is improved mainly because the observed prosody is linguistically unlikely to co-occur with any incorrect word string. Additional improvements, in both perplexity and WER, can be obtained using a semi-factored language model, in which the relationship between prosody and the word sequence is at least partly mediated by syntactic tags. Careful analysis of the relationship between prosody and syntax indicates that syntactic phrase boundaries are the most important cue for prosodic phrase boundary recognition, while part of speech is the most important cue for locating pitch accents, but that neither of these cues is entirely sufficient for either classification task. Experiments to port this system from Radio News to the Switchboard corpus are currently under way, but preliminary results suggest that the prosody of Switchboard is profoundly different from the prosody of Radio News.

1 Introduction

In prosody-dependent speech recognition, acoustic phone models and acoustic prosody models are interdependent, so one cannot be searched without simultaneously searching the other. Our first experiments in prosody-dependent recognition did not explicitly model syntactic structure, but we have discovered that the relationship between word sequence and prosodic tag sequence is most accurately learned by a factored language model with an explicit representation of syntactic class and, if possible, syntactic phrase structure. Thus our systems require an integrated probabilistic model of the relationship among prosodic, syntactic, lexical, and acoustic features, summarized as

$$\hat{W},\hat{P} = \arg\max_{W,P}\max_{S} p(W,S,P,O) \qquad (1)$$

where $O = [\vec{o}_1,\ldots,\vec{o}_T]$ is a sequence of acoustic observations, $W = [w_1,\ldots,w_M]$ is a sequence of words, and each word is tagged with both syntactic information, $S = [s_1,\ldots,s_M]$ and prosodic information $P = [p_1,\ldots,p_M]$. As in most large vocabulary speech recognition systems, Eq. 1 is implemented by way of an intermediate sequence of allophone labels, $Q = [q_1,\ldots,q_L]$, thus

$$p(W,S,P,O) \approx \max_Q p(O|Q)p(Q|W,P) \\ p(W,P|S)p(S) \qquad (2)$$

Enabling technologies for the recognition of prosody include prosody-dependent allophones and prosody-sensitive acoustic observations, discussed in Sec. 2. Enabling technology for the simultaneous recognition of syntax is a factored prosody-dependent language model, with factors representing part of speech and (in a rescoring pass) CFG parse structure, discussed in Sections 3 and 5. The system has been trained and tested using the Radio News Corpus (Ostendorf et al., 1995). The Radio News Corpus is the largest publicly available corpus labeled with the tones and break indices (TOBI) prosodic labeling standard (Beckman and Elam, 1994). The Radio News Corpus was designed for speech synthesis studies. By speech recognition standards, it is an extremely small

corpus (a bit over 3 hours of speech, read by seven professional radio announcers). To our knowledge, no other research group has reported speech recognition word error rate for this corpus, but two studies have reported automatic pitch accent recognition results for this corpus. (Ostendorf and Ross, 1997) achieved 89% accent recognition correctness given known word alignment. (Taylor, 2000) reported 72.7% accent recognition correctness, and 47.7% accent recognition accuracy, based purely on observation of F0 (without lexical sequence information or MFCC observations). Our experiments are not directly comparable to either of these previous studies; like Taylor, we do not assume *a priori* knowledge of word boundary times, but like Ostendorf and Ross, we use word sequence information to aid us in the automatic labeling of prosodic tags.

Current experiments seek to extend our system to the Switchboard corpus. In order to train on Switchboard, it is necessary, first, to both manually and automatically generate TOBI labels for a certain amount of Switchboard data, and second, to develop acoustic and language models of disfluency. Sec. 6 describes our preliminary attempts to transcribe and model the prosody of disfluency in Switchboard.

2 Prosody-Dependent Allophones

In the notation of Eq. 2, recognition of prosody is enabled by the use of prosody-dependent allophone models and prosody-sensitive acoustic observations. Prosody-dependent allophones are similar to the clustered triphones used in standard LVCSR, except that clusters may be defined on the basis of prosodic as well as phonetic context. Each allophone model is a three-state HMM with an explicit duration PMF, and with two observation streams. The first observation stream carries acoustic-phonetic observations (currently MFCCs and energy). The second observation stream carries acoustic-prosodic observations (pitch). Allophone cluster definitions are created separately for the duration PMFs, acoustic phonetic PDFs, and acoustic prosodic PDFs, thus each type of observation is used to distinguish only those context variables with which it is most highly correlated.

Allophone clusters may be defined by any of the following five context variables: left phonetic context, right phonetic context, pitch accent, intonational phrase position, syntactic category. A complete specification of these five context variables may be encoded using the notation shown in Table 1. An allophone is considered to be phrase-final if it is part of the rhyme of the syllable preceding an intonational phrase boundary, and non-final otherwise (Wightman et al., 1992). An allophone is considered to be accented if it is part of the lexically stressed syllable of a word transcribed as containing a pitch accent, and unaccented otherwise. Other prosodic distinctions that we have tested include the distinction between phrase-initial and non-initial allophones, and the distinction between consonants in the onset and coda of an accented syllable (Borys, 2003a; Chen et al., 2004); our best-performing systems implement only the context variables listed in Table 1. Our best-performing systems currently only distinguish the manner class of phonemes to the left and right, and not their place, voicing, or vocalic features (Borys, 2003b; Chen et al., 2004). As place, voicing, and vocalic features have often proven to be useful in other studies of allophonic variation, we suspect that the uselessness of these features in our studies may be an artifact of the relatively small speech corpus that we use to train and test our models. Finally, the "syntactic category" tag may carry any syntactic features that can modify allophone pronunciation without producing a pitch accent or phrase boundary; our current system distinguishes allophones in content words vs. function words.

Table 1: Context variables that may be used to determine an allophone cluster. A fully specified allophone of phoneme `PH` takes the form `L-PH+R_APS`.

Variable	Meaning	Allowed Settings
`L`	Left Phoneme	(vwl, gld, nsl, fric, stop)
`R`	Right Phoneme	(vwl, gld, nsl, fric, stop)
`A`	Accent	(unaccented, accented)
`P`	Phrase	(non-final, final)
`S`	Syntax	(content, function)

Each allophone model is a three-state hidden Markov model with an explicit duration probability mass function (PMF). Modifications to HTK necessary in order to implement an explicit duration PMF are described in (Chen et al., 2004); `diff` files creating the modified functions HDRest, HDERest, HDInit, and HDVite are available at http://www.ifp.uiuc.edu/speech/software/. Parameters of the duration PMF and observation PDFs may be tied independently of one another.

Each state in the allophone model observes two streams of data: an acoustic-phonetic stream, intended to carry information primarily about the shape of the vocal tract, and an acoustic-prosodic stream, intended to carry information primarily about the voice source. The acoustic-prosodic observation stream models a smoothed, nonlinearly transformed pitch frequency, based on the pitch frequency f_0 and probability of voicing (PV) estimated by the `formant` program in Entropic XWAVES. In order to remove pitch doubling and halving errors, we use a method similar to that proposed in (Kompe, 1997): a 3 mixture Gassian classifier is trained on the f_0 data from each utterance, with mixture component means constrained to equal $1/2$, 1, and 2

times the utterance mean pitch. Measured f_0 candidates classified as apparently equal to $2f_0$ or $f_0/2$ are eliminated, as are f_0 measurements with small PVs. Remaining f_0 measurements are normalized and converted to log scale using the formula

$$\hat{f}_0 = \log\left(\frac{f_0}{\mu} + 1\right), \quad (3)$$

where μ is the utterance mean pitch. Eq. 3 is intended to mimic Fujisaki's $\log(f_0/\min f_0)$ parameterization (Fujisaki and Hirose, 1984; Hirai et al., 1997); in our experiments we found that estimates of the mean pitch are less sensitive to pitch tracking errors than estimates of the min f_0, thus we find that Eq. 3 is less sensitive to pitch tracking errors than Fujisaki's parameterization. Frames with missing $\hat{f}_0$ are filled by linearly interpolating $\hat{f}_0$ between available frames, resulting in a smoothed normalized pitch waveform $\tilde{f}_0(t)$. The acoustic-prosodic observation stream models a scalar observation, $Y(t) = g([\tilde{f}_0(t - 20\text{ms}), \ldots, \tilde{f}_0(t + 20\text{ms})])$. The function $g(\cdot)$ is a multilayer perceptron, trained so that $Y(t)$ is an estimate of the *a posteriori* probability that frame t is part of a pitch-accented syllable (Kim et al., in press).

In our best-performing systems, the duration PMF depends on intonational phrase position, and the F0 stream depends on pitch accent. Splitting the duration PMF and the F0 stream seems to be effective, despite the small size of the database, because each of these PDFs requires a very small number of trainable parameters. The duration PMF is stored as a discrete distribution, with 10-15 trainable parameters per state. A Gaussian model of the scalar F0 stream works as well, in our experiments, as a mixture Gaussian model, thus the F0 stream requires only 2 trainable parameters per state. The MFCC stream, by comparison, requires 237 trainable parameters per state for a 3-mixture model of a 39-dimensional acoustic feature vector. Because of the relatively high parameter dimension of the MFCC PDF, all of our attempts to condition the MFCC stream on prosodic context have been stymied by data sparsity problems.

Our approach to prosodic conditioning of the MFCC stream is similar to that proposed in (Ostendorf et al., 1997). Either individual HMM states (as in (Ostendorf et al., 1997)) or entire allophone models are first split into prosody-dependent allophones (as shown in Table 1), then clustered using a standard cross-entropy-based hierarchical clustering algorithm. Two baselines are used: a recognizer composed of clustered prosody-independent triphone models, and a recognizer composed of monophone models. First, we attempted to cluster individual HMM states using the HTK hierarchical clustering routines (HERest, HLStats, and HHEd); embedded reestimation using HERest failed repeatedly to converge, apparently because the training database is just too small. Second, we wrote our own code to cluster entire allophone models using a cross-entropy metric comparable to that used by HTK (Borys, 2003b; Borys, 2003a). In order to guarantee convergence, the clustering algorithm was constrained to generate a pre-specified number of clustered allophone models, regardless of the resulting change in WER. Hierarchical clustering of triphone or allophone models successfully improved the cross-entropy of the test data, but WER of the clustered-allophone recognizer was substantially worse than WER of a 48-monophone baseline recognizer. WER of the monophone recognizer was 24.8%; WER of the prosody-independent clustered triphone model was 36.2%; WER of the prosody-dependent clustered allophone model was 25.2%. Because clustered allophones failed to outperform a monophone model, all results reported in Sec. 4 of this paper will be based on a 48-monophone recognizer, with prosodic splitting only of the duration PMF and the F0 stream.

Although monophones outperform any triphone or allophone model of this database, there are tendencies in the clustering result that support prior phonetic literature in interesting ways. Of the questions selected by the clustering algorithm, slightly more questions concerned intonational phrase position than pitch accent (21% vs. 16%) (Borys, 2003b), in agreement with a number of phonetic studies that suggest important articulatory correlates of intonational phrase boundary (Fougeron and Keating, 1997; Dilley et al., 1996; Cho, 2001). Although vowels were sensitive to all possible prosodic distinctions, consonants were sensitive only to the "lengthening vs. strengthening vs. neutral" three-way distinction proposed by Fougeron and Keating (Fougeron and Keating, 1997): phrase-final consonants ("lengthened") were insensitive to pitch accent, while consonants at the beginning of a phrase-medial accented syllable were grouped together with both accented and unaccented phrase-initial consonants ("strengthened") (Borys, 2003a).

3 Prosody-Dependent Language Models

The relationship between syntax, prosody, and the word string is modeled by a tagged language model. A tagged language model is an estimate of the probability $p(w_m, p_m, s_m|\text{history})$ where w_m is the mth word in the sentence, and p_m and s_m are its prosodic and syntactic tags, respectively. The amount of prosodically labeled data in the English language is not nearly sufficient to create a reliable maximum likelihood estimate of $p(w_m, p_m, s_m|\text{history})$, therefore we have experimented with three methods for estimating the language model probability: a backed-off prosodically-labeled bigram (with no encoding of syntax), and two factored language models.

A prosody-dependent bigram is an estimate of

$p(w_m, p_m|w_{m-1}, p_{m-1})$. The prosodic label p_m carries two types of information: the pitch accent status of word w_m, and the position of w_m within an intonational phrase. There are eight possible settings of p_m: a word may be accented or unaccented; the same word may be phrase-initial, phrase-final, phrase-medial, or it may be a one-word intonational phrase (both phrase-initial and phrase-final). A prosodically tagged word may be encoded in the form `W_AP`, where `W` is the word label, `A` takes the values "a" or "u" (accented or unaccented), and `P` takes the values "i,m,f,o" (initial, medial, final, one-word phrase). The sequence $[p_{m-1}, p_m]$ takes on $|P|^2 = 64$ possible values, so in theory, a prosody-dependent bigram model learns 64 times as many parameters as a prosody-independent bigram model. In practice, most possible combinations of w_m and p_m never occur, so their probabilities are estimated by backing off to 1-gram and 0-gram (uniform) distributions; in our experiments, the actual parameter count of a prosody-dependent bigram model is slightly less than three times that of a prosody-independent bigram.

An empirically superior estimate of the prosody-dependent bigram probability may be trained by explicitly modeling the relationship between the prosodic tag, p_k, and the syntactic tag, s_k (Chen and Hasegawa-Johnson, 2003). The syntactic tag s_k specifies the part of speech of word w_k, and during second-pass decoding (given a complete sentence hypothesis), may also specify the position of word w_k relative to syntactic phrase and clause boundaries. By explicitly modeling syntactic tags, the prosody-dependent bigram probability may be written as

$$p(w_j, p_j|w_i, p_i) = \sum_{s_j s_i} p(w_j, p_j, s_j, s_i|w_i, p_i) \quad (4)$$

$p(w_j, p_j, s_j, s_i|w_i, p_i)$ is proportional to the bigram probability of a syntactically and prosodically tagged vocabulary. This tagged bigram probability may be computed as

$$p(w_j, p_j, s_j, s_i|w_i, p_i) \approx$$
$$p(p_j|s_j, s_i, p_i)p(s_j, s_i|w_j, w_i)p(w_j|w_i, p_i) \quad (5)$$

The approximation in Eq. 5 is valid if we assume that, first, prosody is independent of the word string given knowledge of syntax (reasonable because neither side of the equation has any explicit representation of dialog context), and second, that the syntactic tags are independent of prosody given knowledge of the word string (reasonable except for those cases when prosody may be used to resolve syntactic ambiguity, (Price et al., 1991)). Under these assumptions, the tagged bigram probability factors into three terms. The first term, $p(p_j|s_j, s_i, p_i)$, may be robustly estimated from a relatively small corpus, because the syntactic tagset and the prosodic tagset are both much smaller than the vocabulary. The second term, $p(s_j, s_i|w_j, w_i)$, is the probability that a word sequence (w_i, w_j) implements syntactic tag sequence (s_i, s_j). Computation of this probability is simplified by appropriate choice of the syntactic tagset. During first-pass recognition, the syntactic tag s_i encodes only the part of speech of word w_i. In most cases, the word sequence (w_i, w_j) uniquely determines the POS sequence (s_i, s_j); the few common exceptions can be robustly estimated from a large text database with manual or automatic POS tags. During second-pass recognition, in an N-best rescoring paradigm, it is possible to assume that the recognizer is computing the prosody-dependent and syntax-dependent probability of a complete sentence transcription, $W = [w_1, \ldots, w_M]$. Given a complete transcription, it is possible to compute the maximum likelihood phrase-level parse of the sentence using a context-free grammar, and to augment the syntactic tag s_i with information about the position of the word in its surrounding phrase and clause. Like POS, this new syntactic information may be treated, by the prosody-dependent language model, as information uniquely determined by the hypothesized word sequence (w_i, w_j).

The third term in Eq. 5, $p(w_j|w_i, p_i)$, is a prosody-dependent semi-bigram probability. We have tested two variants of Eq. 5: one in which the probability $p(w_j|w_i, p_i)$ is estimated directly from the Radio News corpus, using backed-off ML estimation, and one in which the probability is estimated using the following approximation:

$$p(w_j|w_i, p_i) = \frac{p(p_i|w_j, w_i)p(w_j|w_i)}{p(p_i|w_i)} \approx$$
$$\frac{\sum_{s_i, s_j} p(p_i|s_i, s_j)p(s_i, s_j|w_i, w_j)p(w_j|w_i)}{\sum_{w_j}\sum_{s_i, s_j} p(p_i|s_i, s_j)p(s_i, s_j|w_i, w_j)p(w_j|w_i)} \quad (6)$$

4 Results

Table 2 describes performance of six different recognizers, each based on 48 monophone HMMs, each composed of an MFCC observation stream (3-mixture Gaussian) and a pitch observation stream (Gaussian), with explicit representation of duration probability density. Each row was created by training the named recognizer on about 90% of the TOBI-transcribed data in the Radio News corpus (six talkers), and testing on the remaining 10% (from the same six talkers). During testing, each recognizer output its best estimate of the complete lexical and prosodic transcription of the utterance. Word error rate was computed by comparing the lexical transcription to a reference using the program HResults, without considering the prosodic transcription; accent and boundary recognition error rates were computed by ignoring the lexical transcription. The system in the first row has no explicit representation of

Table 2: Word error rate (WER), accent error rate (AER), and intonational phrase boundary error rate (BER, in percent) with six different combinations of acoustic model (AM) and language model (LM). PI=prosody independent (baseline), PD=prosody dependent. Accent and boundary error rates of the system with no prosody dependence are at chance.

AM	LM	WER	AER	BER
PI	PI	24.8	44.6	15.6
PD	PI	24.0	45.9	15.0
PI	PD Bigram	24.3	23.1	14.5
PD	PD Bigram	23.4	20.3	14.3
PD	PD Semi-factored	21.7	20.3	14.2
PD	PD Factored	22.9	19.7	13.4

prosody. Accent and boundary recognition error rates of the first system are at chance for this database: 45% of words in this database are unaccented (55% are accented), and 16% are phrase-final. In the second system, the F0 stream is accent-dependent and the duration PMF is phrase-position dependent; all systems in this table use a prosody-independent MFCC stream. The third system uses a prosody-dependent bigram language model with no model of the acoustic correlates of prosody. The fourth system uses a prosody-dependent bigram, plus explicit models of accent-dependent pitch variation and phrase-final lengthening. The fifth system uses a semi-factored language model, meaning that $p(w_j, p_j|w_i, p_i)$ is factored, but $p(w_j|w_i, p_i)$ is not (Eq. 5 is used, but not Eq. 6). The last system uses both Eq. 5 and 6.

The results of Table 2 indicate that word error rate is only significantly improved if a prosody-dependent acoustic model and a prosody-dependent language model are combined. Prosody-dependent language modeling, alone, is sufficient for better-than-chance recognition of accents and boundaries; a prosody-dependent acoustic model, alone, is insufficient for any type of gain. Chen and Hasegawa-Johnson (Chen and Hasegawa-Johnson, 2004) have presented a formal, information-theoretic hypothesis explaining the necessity of simultaneous prosody-dependent language modeling and acoustic modeling. The core of the argument is the observation that word error rate is improved only if the observed prosody (the prosody that maximizes the acoustic observation PDF) is linguistically unlikely to co-occur with any incorrect word string.

The last two rows of Table 2 present results obtained using the semi-factored and factored bigram language models. The word perplexities of the bigram, semi-factored, and factored language models, using the same test corpus as in Table 2, are 60, 54, and 47, respectively. The semi-factored model has significantly lower WER than the baseline bigram (21.7% vs. 23.4%), but not significantly lower boundary error rate (14.2% vs. 14.3%) or accent error rate (20.3% vs. 20.3%). The factored model has significantly improved boundary recognition error (13.4% vs. 14.3%), but not significantly improved WER (22.9% vs. 23.4%).

Table 3: Accent error rate (AER) and boundary error rate (BER) of five machine learning algorithms in the task of automatic prosodic transcription of the Radio News corpus based on word sequence information. NN=Neural Network. From (Cohen, 2004).

Features: Word-based POS		
Learning Algorithm	AER	BER
None (Chance)	42.6	19.1
C4.5 Rules	18.1	12.1
SLIPPER	18.4	11.5
QUEST Univariate		12.1
Features: Full Syntactic Parse		
C4.5 Rules	17.3	11.2
SLIPPER	17.7	10.2
QUEST Univariate	17.5	10.6
QUEST Linear	17.4	11.0
NN, All Features	17.1	10.8
NN, Category Features	16.9	10.4

5 Prediction of Prosody from Syntax

Table 2 demonstrates that prosody is most useful when its acoustic and word sequence correlates are jointly modeled, and that, for the purpose of modeling prosody, syntactically inspired language models significantly outperform a baseline bigram model. In order to better understand the relationship between prosody and syntax, Cohen (Cohen, 2004) performed a series of experiments in which automatic syntactic parsers, tree-based learners, and neural networks were used to predict the prosodic tags on each word in the Radio News corpus. Nine machine learning algorithms were tested, using seven different syntactic feature sets, for prediction of two prosodic tag variables. Table 3 presents results from several representative experiments, including the most successful.

All classifiers in Table 3 used a two-stage classification algorithm: word sequence was first automatically parsed to produce syntactic tags, and syntactic tags were then classified in order to determine prosodic tags. Two types of binary prosodic tags were estimated: accent recognition marked the target word as either accented or unaccented, while boundary recognition marked the target word as either intonational-phrase-initial or non-initial.

Two types of syntactic parsers were used. The top half of the table, marked "Word-Based POS," describes ac-

cent recognition experiments using part of speech (POS) information generated by the Roth-Zelenko (RZ) shallow parsing algorithm (Roth and Zelenko, 1998), and prosodic boundary recognition experiments using the RZ algorithm plus seven syntactic phrase boundary features. Each prosodic tag is computed based on observation of the POS of three consecutive words (the target word plus two prior words). Each POS tag, in turn, is computed by the RZ algorithm based on lexical features in a five word window, thus the total system computes prosodic tags of one word based on lexical features of 3+5-1=7 consecutive words. Recognition of intonational phrase boundary based only on POS information was found to be quite poor, therefore boundary recognition results in this half of the table also use a small set of full-parse information: seven features labeling the type of the syntactic phrase boundary beginning on the target word, as determined by the Charniak parser. The columns marked "Full Syntactic Parse" use both POS and phrasal parse information generated by Charniak's parser (Charniak, 1994). Charniak's parser computes the maximum likelihood parse of an entire breath group using a stochastic context-free grammar, including opening and closing of every phrase, clause, and fragment, and the part of speech of every word. Prosodic taggers based on full-parse features observed POS in a four-word window, and parse features in a two-word window (the target word and the word to its left). Parse features of a word include indicator features marking the types of phrases that open or close with the given word, as well as two integer features counting the number of phrases, of any type, that the word opens or closes. Two types of indicator features were tested: the "all features" condition used separate indicator features for every type of phrase or clause defined by the Charniak parser, while the "category features" used indicator features to mark the onset and offset of heuristically designed categories. All learners except the neural network had better performance in the "all features" condition, thus results for the "category features" are only shown for the neural network.

Five learners were tested. The neural network was a sigmoidal feedforward network trained using error backpropagation. SLIPPER is a boosting algorithm based on the RIPPER rule learner (Cohen, 1995). C4.5 and QUEST are tree-based learners (Quinlan, 1993; Loh and Shih, 1997). Trees learned by the C4.5 algorithm were generally found to have better test-corpus accuracy if the tree was first post-processed in order to generate a series of rules; the resulting rules were also considerably more human-legible than the raw learned trees. QUEST was tested using either univariate nodes or "linear" nodes; the linear-node configuration implements a linear discriminant combination of all features at each node in the tree. Rulesets learned by univariate QUEST were generally more concise and more legible, to a human expert, than those learned by any other algorithm.

In some cases, different learners discovered quite different patterns of information. The C4.5 learner classifies pitch accent on the basis of both phrase and POS information, if both are available: for example, with certain exceptions, words are marked as accented if they close a subordinate clause but not a prepositional phrase. The QUEST learner, on the other hand, determines pitch accent using rules that consider only the POS of the current word and next word. The first QUEST rule places a pitch accent on every noun, adjective, gerund, or participle, regardless of context. It must be noted, however, that even though the QUEST learner uses only POS to determine pitch accent, the QUEST classifier learned using a full CFG parse outperforms the classifier learned using local word-based POS tags. Apparently, POS tags generated by the CFG parser are more useful, for the purpose of prosody recognition, than POS tags generated by a local word-sequence-based tagger.

The best predictor of intonational phrase boundary is the simultaneous closure of more than one syntactic phrase. If syntactic parse information is unavailable, the best predictor is the presence of an audible breath; in this corpus, breath can be pretty reliably labeled based on duration of the pause. After these cues, the next several levels in both trees are primarily occupied by POS features. For example, an intonational phrase boundary is likely after a noun or interjection, or before an auxiliary, coordinating conjunction, modal verb, or the word "to."

6 Switchboard

Our current research seeks to extend these results to spontaneous speech. Using the neural network classifier whose performance is listed in Table 3, we have tagged syntactically predicted accent and intonational phrase boundary positions in the Switchboard conversational telephone speech corpus (Godfrey et al., 1992). In order to test these results, and in order to learn about the differences between conversational speech and read speech, we have started to manually transcribe the prosody and disfluency segments in the WS97 subset of Switchboard (Greenberg and Hitchcock, 2001; Chavarria et al., 2004).

Preliminary results from this corpus indicate that statistical models trained to represent the prosody of Radio News speech are unable to predict the prosody of Switchboard speech. The models listed in Table 3 do not predict prosodic phrase boundary position at rates better than chance. Pitch accent is predicted with an accuracy better than chance, but still insufficient to be of any use for the clustering of allophone HMMs, thus when clustering experiments are repeated on the Switchboard corpus, error rates of the prosody-dependent and prosody-independent

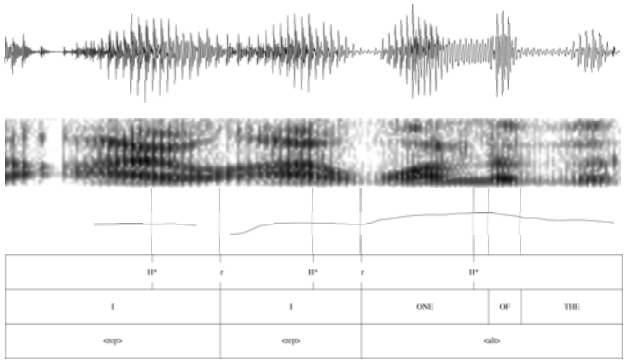

Figure 1: Transcription of prosody and disfluencies in the phrase "I, I, one of the..."

systems are identical.

Manual transcription suggests many reasons why the syntactic and acoustic correlates of prosody in the Switchboard corpus may be significantly different from their correlates in Radio News. First, few Switchboard utterances contain complete, well-formed sentences. Second, Radio News speech is characterized by clear F0 markers for all kinds of pitch accent and phrase boundary tones, while F0 contours extracted from the Switchboard corpus are comparatively monotone. Comparison of L- (intermediate phrase boundary tones) and L-L% (intonational phrase boundary tones) on the Switchboard corpus discovered that phrase-final lengthening is the only reliable acoustic correlate of this distinction; F0 correlates seem not to reliably mark this distinction in Switchboard (Chavarria et al., 2004).

Disfluency is the third reason that Switchboard prosody is unlike Radio News prosody. Although we began transcribing Switchboard with the goal of only annotating prosody, we discovered almost immediately that it is impossible to annotate prosodic phrase boundaries in Switchboard without devising some sort of annotation for disfluencies. We have adopted the annotation system of Heeman and Allen (Heeman and Allen, 1999), according to which the words being corrected are called the "reparandum" or REP, the correction is called the "alteration" (ALT), and filled pauses or meta-dialog between REP and ALT are called the "edit" (EDT).

Fig. 1 shows a disfluency with a double reparandum: "I, I, one of the things I..." The first reparandum is repeated, then finally replaced by the alteration. Fig. 1 shows two characteristics of disfluency that have not been extensively studied. First, both of the reparanda end in glottalization, clearly visible in the form of extremely low-frequency or low-amplitude glottal excitation. Second, the prosody of the reparandum is "mimicked" in the alteration, despite dramatically different lexical content. In this corpus, words in the reparandum or alteration of a disfluency are as likely to bear a pitch accent as any other words in the sentence: 40%, compared to 39.9% of all words. About two thirds of the accented words in the reparandum are replaced by accented words in the alteration (10/16); about two thirds of unaccented words are replaced by unaccented words (15/21). Repetition of prosody is perceptually salient: some listeners report that the alteration "mimics" the intonational contour of the reparandum.

Disfluency is common in Switchboard. Of 1100 words we have transcribed, 40 are part of a reparandum, 37 are filled pauses, and 41 are part of an alteration, thus 10% of the words we have transcribed are part of a disfluency. This estimate is higher than most published estimates, perhaps because we include all words that are part of the reparandum or alteration, but most published studies estimate that at least 5% of the words in Switchboard are part of a disfluency (e.g., (Shriberg, 2001)). Any complete description of the prosody of Switchboard will necessarily include, as one component, a theory about the prosody of disfluency.

7 Conclusions

This paper has reviewed results from a number of experimental systems that simultaneously recognize the prosodic and lexical transcriptions of an utterance. It has been demonstrated, first, that prosody-dependent allophone modeling can reduce the word error rate of a speech recognizer, but that reliable WER reductions depend on the simultaneous use of both a prosody-dependent acoustic model and a prosody-dependent language model. Additional improvements, in both perplexity and WER, can be obtained using a semi-factored language model, in which the relationship between prosody and the word sequence is at least partly mediated by syntactic tags. Careful analysis of the relationship between prosody and syntax indicates that syntactic phrase boundaries are the most important cue for prosodic phrase boundary recognition, while part of speech is the most important cue for locating pitch accents, but that neither of these cues is entirely sufficient for either classification task. Even if a pitch accent recognizer completely ignores phrase information (as does the QUEST learner), its error rate can be reduced by deriving POS information from a complete CFG parse of the sentence, rather than from a local lexical-feature-based classifier.

The prosody of conversational telephone speech is significantly different, in important ways, from the prosody of Radio News speech. Preliminary results suggest that important differences include the use of incomplete sentences, relatively greater use of duration to cue prosody (and correspondingly less use of pitch), and, perhaps most importantly, the frequent occurrence of disfluency.

References

M. E. Beckman and G. A. Elam. 1994. Guidelines for ToBI labelling. Technical report, Ohio State University. http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/singer_tobi.html.

Sarah Borys. 2003a. The importance of prosodic factors in phoneme modeling with applications to speech recognition. In *HLT/NAACL student session*, Edmonton.

Sarah Borys. 2003b. Recognition of prosodic factors and detection of landmarks for automatic speech recognition. Bachelor's thesis, University of Illinois at Urbana-Champaign.

Eugene Charniak. 1994. *Statistical Language Learning*. MIT Press, Cambridge, MA.

Sandra Chavarria, Taejin Yoon, Jennifer Cole, and Mark Hasegawa-Johnson. 2004. Acoustic differentiation of ip and IP boundary levels: Comparison of L- and L-L% in the switchboard corpus. In *ISCA Internat. Conf. Speech Prosody*, Nara, Japan.

Ken Chen and Mark Hasegawa-Johnson. 2003. Improving the robustness of prosody dependent language modeling based on prosody syntax cross-correlation. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Ken Chen and Mark Hasegawa-Johnson. 2004. How prosody improves word recognition. In *ISCA Internat. Conf. Speech Prosody*, Nara, Japan.

Ken Chen, Mark Hasegawa-Johnson, Aaron Cohen, Sarah Borys, Sung-Suk Kim, Jennifer Cole, and Jeung-Yoon Choi. 2004. Prosody dependent speech recognition on radio news. (in review).

T. Cho. 2001. *Effects of Prosody on Articulation in English*. Ph.D. thesis, UCLA.

William W. Cohen. 1995. Fast effective rule induction. In *Proc. International Conference on Machine Learning*.

Aaron Cohen. 2004. A survey of machine learning methods for predicting prosody in radio speech. Master's thesis, University of Illinois at Urbana-Champaign.

Laura Dilley, Stefanie Shattuck-Hufnagel, and Mari Ostendorf. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *J. of Phonetics*, 24:423–444.

Cecile Fougeron and Patricia A. Keating. 1997. Articulatory strengthening at edges of prosodic domains. *J. Acoust. Soc. Am*, 101(6):3728–3740.

H. Fujisaki and K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentence of Japanese. *J. Acoust. Soc. Japan*, 5(4):233–242.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proc. ICASSP*, pages 517–520.

Steven Greenberg and Leah Hitchcock. 2001. Stress-accent and vowel quality in the Switchboard corpus. In *NIST Large Vocabulary Continuous Speech Recognition Workshop*, Linthicum Heights, MD, May.

Peter A. Heeman and James F. Allen. 1999. Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4).

Toshio Hirai, Naoto Iwahashi, Norio Higuchi, and Yoshinori Sagisaka. 1997. Automatic extraction of f_0 control rules using statistical analysis. In Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis*, pages 333–346. Springer-Verlag, New York.

Sung-Suk Kim, Mark Hasegawa-Johnson, and Ken Chen. (in press). Automatic recognition of pitch movements using multi-layer perceptron and time-delay recursive neural network. *IEEE Signal Processing Letters*.

R. Kompe. 1997. *Prosody in Speech Understanding Systems*. Springer-Verlag.

W.-Y. Loh and Y.-S. Shih. 1997. Split selection methods for classification trees. *Statistica Sinica*, 7:815–840.

M. Ostendorf and K. Ross. 1997. A multi-level model for recognition of intonation labels. In *Computing prosody: computational models for processing spontaneous speech*. Springer-Verlag New York, Inc.

M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel. 1995. *The Boston University Radio News Corpus*. Linguistic Data Consortium.

M. Ostendorf, B. Byrne, M. Fink, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfield. 1997. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode: Final report. Technical Report WS96, Johns Hopkins University Center for Language and Speech Processing.

P.J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. 1991. The use of prosody in syntactic disambiguation. *J. Acoust. Soc. Am*, 90(6):2956–2970, Dec.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufman, Boston.

Dan Roth and Dmitry Zelenko. 1998. Part of speech tagging using a network of linear separators. In *COLING-ACL*.

Elizabeth Shriberg. 2001. To 'errr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–164.

Paul Taylor. 2000. Analysis and synthesis of intonation using the Tilt model. *J. Acoust. Soc. Am*, 107(3):1697–1714.

Colin Wightman, Stefanie Shattuck-Hufnagel, and Mari Ostendorf andPatti Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *J. Acoust. Soc. Am*, 91(3):1707–1717, March.