N-Best Rescoring Based on Pitch-accent Patterns

Je Hun Jeon[1] **Wen Wang**[2] **Yang Liu**[1]

[1]Department of Computer Science, The University of Texas at Dallas, USA
[2]Speech Technology and Research Laboratory, SRI International, USA
{jhjeon,yangl}@hlt.utdallas.edu, wwang@speech.sri.com

Abstract

In this paper, we adopt an n-best rescoring scheme using pitch-accent patterns to improve automatic speech recognition (ASR) performance. The pitch-accent model is decoupled from the main ASR system, thus allowing us to develop it independently. N-best hypotheses from recognizers are rescored by additional scores that measure the correlation of the pitch-accent patterns between the acoustic signal and lexical cues. To test the robustness of our algorithm, we use two different data sets and recognition setups: the first one is English radio news data that has pitch accent labels, but the recognizer is trained from a small amount of data and has high error rate; the second one is English broadcast news data using a state-of-the-art SRI recognizer. Our experimental results demonstrate that our approach is able to reduce word error rate relatively by about 3%. This gain is consistent across the two different tests, showing promising future directions of incorporating prosodic information to improve speech recognition.

1 Introduction

Prosody refers to the suprasegmental features of natural speech, such as rhythm and intonation, since it normally extends over more than one phoneme segment. Speakers use prosody to convey paralinguistic information such as emphasis, intention, attitude, and emotion. Humans listening to speech with natural prosody are able to understand the content with low cognitive load and high accuracy. However, most modern ASR systems only use an acoustic model and a language model. Acoustic information in ASR is represented by spectral features that are usually extracted over a window length of a few tens of milliseconds. They miss useful information contained in the prosody of the speech that may help recognition.

Recently a lot of research has been done in automatic annotation of prosodic events (Wightman and Ostendorf, 1994; Sridhar et al., 2008; Ananthakrishnan and Narayanan, 2008; Jeon and Liu, 2009). They used acoustic and lexical-syntactic cues to annotate prosodic events with a variety of machine learning approaches and achieved good performance. There are also many studies using prosodic information for various spoken language understanding tasks. However, research using prosodic knowledge for speech recognition is still quite limited. In this study, we investigate leveraging prosodic information for recognition in an n-best rescoring framework.

Previous studies showed that prosodic events, such as pitch-accent, are closely related with acoustic prosodic cues and lexical structure of utterance. The pitch-accent pattern given acoustic signal is strongly correlated with lexical items, such as syllable identity and canonical stress pattern. Therefore as a first study, we focus on pitch-accent in this paper. We develop two separate pitch-accent detection models, using acoustic (observation model) and lexical information (expectation model) respectively, and propose a scoring method for the correlation of pitch-accent patterns between the two models for recognition hypotheses. The n-best list is rescored using the pitch-accent matching scores

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 732–741,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

combined with the other scores from the ASR system (acoustic and language model scores). We show that our method yields a word error rate (WER) reduction of about 3.64% and 2.07% relatively on two baseline ASR systems, one being a state-of-the-art recognizer for the broadcast news domain. The fact that it holds across different baseline systems suggests the possibility that prosody can be used to help improve speech recognition performance.

The remainder of this paper is organized as follows. In the next section, we review previous work briefly. Section 3 explains the models and features for pitch-accent detection. We provide details of our n-best rescoring approach in Section 4. Section 5 describes our corpus and baseline ASR setup. Section 6 presents our experiments and results. The last section gives a brief summary along with future directions.

2 Previous Work

Prosody is of interest to speech researchers because it plays an important role in comprehension of spoken language by human listeners. The use of prosody in speech understanding applications has been quite extensive. A variety of applications have been explored, such as sentence and topic segmentation (Shriberg et al., 2000; Rosenberg and Hirschberg, 2006), word error detection (Litman et al., 2000), dialog act detection (Sridhar et al., 2009), speaker recognition (Shriberg et al., 2005), and emotion recognition (Benus et al., 2007), just to name a few.

Incorporating prosodic knowledge is expected to improve the performance of speech recognition. However, how to effectively integrate prosody within the traditional ASR framework is a difficult problem, since prosodic features are not well defined and they come from a longer region, which is different from spectral features used in current ASR systems. Various research has been conducted trying to incorporate prosodic information in ASR. One way is to directly integrate prosodic features into the ASR framework (Vergyri et al., 2003; Ostendorf et al., 2003; Chen and Hasegawa-Johnson, 2006). Such efforts include prosody dependent acoustic and pronunciation model (allophones were distinguished according to different prosodic phenomenon), language model (words were augmented by prosody events), and duration modeling (different prosodic events were modeled separately and combined with conventional HMM). This kind of integration has advantages in that spectral and prosodic features are more tightly coupled and jointly modeled. Alternatively, prosody was modeled independently from the acoustic and language models of ASR and used to rescore recognition hypotheses in the second pass. This approach makes it possible to independently model and optimize the prosodic knowledge and to combine with ASR hypotheses without any modification of the conventional ASR modules. In order to improve the rescoring performance, various prosodic knowledge was studied. (Ananthakrishnan and Narayanan, 2007) used acoustic pitch-accent pattern and its sequential information given lexical cues to rescore n-best hypotheses. (Kalinli and Narayanan, 2009) used acoustic prosodic cues such as pitch and duration along with other knowledge to choose a proper word among several candidates in confusion networks. Prosodic boundaries based on acoustic cues were used in (Szaszak and Vicsi, 2007).

We take a similar approach in this study as the second approach above in that we develop prosodic models separately and use them in a rescoring framework. Our proposed method differs from previous work in the way that the prosody model is used to help ASR. In our approach, we explicitly model the symbolic prosodic events based on acoustic and lexical information. We then capture the correlation of pitch-accent patterns between the two different cues, and use that to improve recognition performance in an n-best rescoring paradigm.

3 Prosodic Model

Among all the prosodic events, we use only pitch-accent pattern in this study, because previous studies have shown that acoustic pitch-accent is strongly correlated with lexical items, such as canonical stress pattern and syllable identity that can be easily acquired from the output of conventional ASR and pronunciation dictionary. We treat pitch-accent detection as a binary classification task, that is, a classifier is used to determine whether the base unit is prominent or not. Since pitch-accent is usually carried by syllables, we use syllables as our units, and the syllable definition of each word is based on CMU pronunciation dictionary which has lexical stress and syllable boundary marks (Bartlett et al., 2009). We separately develop acoustic-prosodic and lexical-prosodic models and use the correlation between the two models for each syllable to rescore the n-best hypotheses of baseline ASR systems.

3.1 Acoustic-prosodic Features

Similar to most previous work, the prosodic features we use include pitch, energy, and duration. We also add delta features of pitch and energy. Duration information for syllables is derived from the speech waveform and phone-level forced alignment of the transcriptions. In order to reduce the effect by both inter-speaker and intra-speaker variation, both pitch and energy values are normalized (z-value) with utterance specific means and variances. For pitch, energy, and their delta values, we apply several categories of 12 functions to generate derived features.

- Statistics (7): minimum, maximum, range, mean, standard deviation, skewness and kurtosis value. These are used widely in prosodic event detection and emotion detection.
- Contour (5): This is approximated by taking 5 leading terms in the Legendre polynomial expansion. The approximation of the contour using the Legendre polynomial expansion has been successfully applied in quantitative phonetics (Grabe et al., 2003) and in engineering applications (Dehak et al., 2007). Each term models a particular aspect of the contour, such as the slope, and information about the curvature.

We use 6 duration features, that is, raw, normalized, and relative durations (ms) of the syllable and vowel. Normalization (z-value) is performed based on statistics for each syllable and vowel. The relative value is the difference between the normalized current duration and the following one.

In the above description, we assumed that the event of a syllable is only dependent on its observations, and did not consider contextual effect. To alleviate this restriction, we expand the features by incorporating information about the neighboring syllables. Based on the study in (Jeon and Liu, 2010) that evaluated using left and right contexts, we choose to use one previous and one following context in the features. The total number of features used in this study is 162.

3.2 Lexical-prosodic Features

There is a very strong correlation between pitch-accent in an utterance and its lexical information. Previous studies have shown that the lexical features perform well for pitch-accent prediction. The detailed features for training the lexical-prosodic model are as follows.

- Syllable identity: We kept syllables that appear more than 5 times in the training corpus. The other syllables that occur less are collapsed into one syllable representation.
- Vowel phone identity: We used vowel phone identity as a feature.
- Lexical stress: This is a binary feature to represent if the syllable corresponds to a lexical stress based on the pronunciation dictionary.
- Boundary information: This is a binary feature to indicate if there is a word boundary before the syllable.

For lexical features, based on the study in (Jeon and Liu, 2010), we added two previous and two following contexts in the final features.

3.3 Prosodic Model Training

We choose to use a support vector machine (SVM) classifier[1] for the prosodic model based on previous work on prosody labeling study in (Jeon and Liu, 2010). We use RBF kernel for the acoustic model, and 3-order polynomial kernel for the lexical model.

In our experiments, we investigate two kinds of training methods for prosodic modeling. The first one is a supervised method where models are trained using all the labeled data. The second is a semi-supervised method using co-training algorithm (Blum and Mitchell, 1998), described in Algorithm 1. Given a set L of labeled data and a set U of unlabeled data with two views, it then iterates in the

[1] LIBSVM – A Library for Support Vector Machines, location: http://www.csie.ntu.edu.tw/˜cjlin/libsvm/

Algorithm 1 Co-training algorithm.

```
Given:
  - $L$: labeled examples; $U$: unlabeled examples
  - there are two views $V_1$ and $V_2$ on an example $x$
Initialize:
  - $L_1$=$L$, samples used to train classifiers $h_1$
  - $L_2$=$L$, samples used to train classifiers $h_2$
Loop for k iterations
  - create a small pool $U$´ choosing from $U$
  - use $V_1(L_1)$ to train classifier $h_1$
    and $V_2(L_2)$ to train classifier $h_2$
  - let $h_1$ label/select examples $D_{h_1}$ from $U$´
  - let $h_2$ label/select examples $D_{h_2}$ from $U$´
  - add self-labeled examples $D_{h_1}$ to $L_2$
    and $D_{h_2}$ to $L_1$
  - remove $D_{h_1}$ and $D_{h_2}$ from U
```

following procedure. The algorithm first creates a smaller pool U' containing unlabeled data from U. It uses L_i ($i = 1, 2$) to train two distinct classifiers: the acoustic classifier h_1, and the lexical classifier h_2. We use function V_i ($i = 1, 2$) to represent that only a single view is used for training h_1 or h_2. These two classifiers are used to make predictions for the unlabeled set U', and only when they agree on the prediction for a sample, their predicted class is used as the label for this sample. Then among these self-labeled samples, the most confident ones by one classifier are added to the data set L_i for training the other classifier. This iteration continues until reaching the defined number of iterations. In our experiment, the size of the pool U´ is 5 times of the size of training data L_i, and the size of the added self-labeled example set, D_{h_i}, is 5% of L_i. For the newly selected D_{h_i}, the distribution of the positive and negative examples is the same as that of the training data L_i.

This co-training method is expected to cope with two problems in prosodic model training. The first problem is the different decision patterns between the two classifiers: the acoustic model has relatively higher precision, while the lexical model has relatively higher recall. The goal of the co-training algorithm is to learn from the difference of each classifier, thus it can improve the performance as well as reduce the mismatch of two classifiers. The second problem is the mismatch of data used for model training and testing, which often results in system performance degradation. Using co-training, we can use the unlabeled data from the domain that matches the test data, adapting the model towards test domain.

4 N-Best Rescoring Scheme

In order to leverage prosodic information for better speech recognition performance, we augment the standard ASR equation to include prosodic information as following:

$$\begin{aligned} \hat{W} &= \arg\max_{W} p(W|A_s, A_p) \\ &= \arg\max_{W} p(A_s, A_p|W)p(W) \end{aligned} \quad (1)$$

where A_s and A_p represent acoustic-spectral features and acoustic-prosodic features. We can further assume that spectral and prosodic features are conditionally independent given a word sequence W, therefore, Equation 1 can be rewritten as following:

$$\hat{W} \approx \arg\max_{W} p(A_s|W)p(W)p(A_p|W) \quad (2)$$

The first two terms stand for the acoustic and language models in the original ASR system, and the last term means the prosody model we introduce. Instead of using the prosodic model in the first pass decoding, we use it to rescore n-best candidates from a speech recognizer. This allows us to train the prosody models independently and better optimize the models.

For $p(A_p|W)$, the prosody score for a word sequence W, in this work we propose a method to estimate it, also represented as $score_{W-prosody}(W)$. The idea of scoring the prosody patterns is that there is some expectation of pitch-accent patterns given the lexical sequence (W), and the acoustic pitch-accent should match with this expectation. For instance, in the case of a prominent syllable, both acoustic and lexical evidence show pitch-accent, and vice versa. In order to maximize the agreement between the two sources, we measure how good the acoustic pitch-accent in speech signal matches the given lexical cues. For each syllable S_i in the n-best list, we use acoustic-prosodic cues (a_i) to estimate the posterior probability that the syllable is prominent (P), $p(P|a_i)$. Similarly, we use lexical cues (l_i)

to determine the syllable's pitch-accent probability $p(P|l_i)$. Then the prosody score for a syllable S_i is estimated by the match of the pitch-accent patterns between acoustic and lexical information using the difference of the posteriors from the two models:

$$score_{S-prosody}(S_i) \approx 1- \mid p(P|a_i) - p(P|l_i) \mid \quad (3)$$

Furthermore, we take into account the effect due to varying durations for different syllables. We notice that syllables without pitch-accent have much shorter duration than the prominent ones, and the prosody scores for the short syllables tend to be high. This means that if a syllable is split into two consecutive non-prominent syllables, the agreement score may be higher than a long prominent syllable. Therefore, we introduce a weighting factor based on syllable duration ($dur(i)$). For a candidate word sequence (*W*) consisting of n syllables, its prosodic score is the sum of the prosodic scores for all the syllables in it weighted by their duration (measured using milliseconds), that is:

$$score_{W-prosody}(W) \approx \sum_{i=1}^{n} log(score_{S-prosody}(S_i)) \cdot dur(i) \quad (4)$$

We then combine this prosody score with the original acoustic and language model likelihood ($P(A_s|W)$ and $P(W)$ in Equation 2). In practice, we need to weight them differently, therefore, the combined score for a hypothesis W is:

$$Score(W) = \lambda \cdot score_{W-prosody}(W) + score_{ASR}(W) \quad (5)$$

where $score_{ASR}(W)$ is generated by ASR systems (composed of acoustic and language model scores) and λ is optimized using held out data.

5 Data and Baseline Systems

Our experiments are carried out using two different data sets and two different recognition systems as well in order to test the robustness of our proposed method.

The first data set is the Boston University Radio News Corpus (BU) (Ostendorf et al., 1995), which consists of broadcast news style read speech. The BU corpus has about 3 hours of read speech from 7 speakers (3 female, 4 male). Part of the data has been labeled with ToBI-style prosodic annotations. In fact, the reason that we use this corpus, instead of other corpora typically used for ASR experiments, is because of its prosodic labels. We divided the entire data corpus into a training set and a test set. There was no speaker overlap between training and test sets. The training set has 2 female speakers (*f2* and *f3*) and 3 male ones (*m2*, *m3*, *m4*). The test set is from the other two speakers (*f1* and *m1*). We use 200 utterances for the recognition experiments. Each utterance in BU corpus consists of more than one sentences, so we segmented each utterance based on pause, resulting in a total number of 713 segments for testing. We divided the test set roughly equally into two sets, and used one for parameter tuning and the other for rescoring test. The recognizer used for this data set was based on Sphinx-3[2]. The context-dependent triphone acoustic models with 32 Gaussian mixtures were trained using the training partition of the BU corpus described above, together with the broadcast new data. A standard back-off trigram language model with Kneser-Ney smoothing was trained using the combined text from the training partition of the BU, Wall Street Journal data, and part of Gigaword corpus. The vocabulary size was about 10K words and the out-of-vocabulary (OOV) rate on the test set was 2.1%.

The second data set is from broadcast news (BN) speech used in the GALE program. The recognition test set contains 1,001 utterances. The n-best hypotheses for this data set are generated by a state-of-the-art SRI speech recognizer, developed for broadcast news speech (Stolcke et al., 2006; Zheng et al., 2007). This system yields much better performance than the first one. We also divided the test set roughly equally into two sets for parameter tuning and testing. From the data used for training the speech recognizer, we randomly selected 5.7 hours of speech (4,234 utterances) for the co-training algorithm for the prosodic models.

For prosodic models, we used a simple binary representation of pitch-accent in the form of presence versus absence. The reference labels are de-

[2] CMU Sphinx - Speech Recognition Toolkit, location: http://www.speech.cs.cmu.edu/sphinx/tutorial.html

rived from the ToBI annotation in the BU corpus, and the ratio of pitch-accented syllables is about 34%. Acoustic-prosodic and lexical-prosodic models were separately developed using the features described in Section 3. Feature extraction was performed at the syllable level from force-aligned data. For the supervised approach, we used those utterances in the training data partition with ToBI labels in the BU corpus (245 utterances, 14,767 syllables). For co-training, the labeled data from BU corpus is used as initial training, and the other unlabeled data from BU and BN are used as unlabeled data.

6 Experimental Results

6.1 Pitch-accent Detection

First we evaluate the performance of our acoustic-prosodic and lexical-prosodic models for pitch-accent detection. For rescoring, not only the accuracies of the two individual prosodic models are important, but also the pitch-accent agreement score between the two models (as shown in Equation 3) is critical, therefore, we present results using these two metrics. Table 1 shows the accuracy of each model for pitch-accent detection, and also the average prosody score of the two models (i.e., Equation 3) for positive and negative classes (using reference labels). These results are based on the BU labeled data in the test set. To compare our pitch accent detection performance with previous work, we include the result of (Jeon and Liu, 2009) as a reference. Compared to previous work, the acoustic model achieved similar performance, while the performance of lexical model is a bit lower. The lower performance of lexical model is mainly because we do not use part-of-speech (POS) information in the features, since we want to only use the word output from the ASR system (without additional POS tagging).

As shown in Table 1, when using the co-training algorithm, as described in Section 3.3, the overall accuracies improve slightly and therefore the prosody score is also increased. We expect this improved model will be more beneficial for rescoring.

6.2 N-Best Rescoring

For the rescoring experiment, we use 100-best hypotheses from the two different ASR systems, as described in Section 5. We apply the acoustic and lexical prosodic models to each hypothesis to obtain its prosody score, and combine it with ASR scores to find the top hypothesis. The weights were optimized using one test set and applied to the other. We report the average result of the two testings.

	Accuracy(%)		**Prosody score**	
	Acoustic	Lexical	Pos	Neg
Supervised	83.97	84.48	0.747	0.852
Co-training	84.54	84.99	0.771	0.867
Reference	83.53	87.92	-	-

Table 1: Pitch accent detection results: performance of individual acoustic and lexical models, and the agreement between the two models (i.e., prosody score for a syllable, Equation 3) for positive and negative classes. Also shown is the reference result for pitch accent detection from Jeon and Liu (2009).

Table 2 shows the rescoring results using the first recognition system on BU data, which was trained with a relatively small amount of data. The 1-best baseline uses the first hypothesis that has the best ASR score. The oracle result is from the best hypothesis that gives the lowest WER by comparing all the candidates to the reference transcript. We used two prosodic models as described in Section 3.3. The first one is the base prosodic model using supervised training (*S-model*). The second is the prosodic model with the co-training algorithm (*C-model*). For these rescoring experiments, we tuned λ (in Equation 5) when combining the ASR acoustic and language model scores with the additional prosody score. The value in parenthesis in Table 2 means the relative WER reduction when compared to the baseline result. We show the WER results for both the development and the test set.

As shown in Table 2, we observe performance improvement using our rescoring method. Using the base *S-model* yields reasonable improvement, and *C-model* further reduces WER. Even though the prosodic event detection performance of these two prosodic models is similar, the improved prosody score between the acoustic and lexical prosodic models using co-training helps rescoring. After rescoring using prosodic knowledge, the WER is reduced by 0.82% (3.64% relative). Furthermore, we notice that the difference between development and

		WER (%)
1-best baseline		22.64
S-model	Dev	21.93 (3.11%)
	Test	22.10 (2.39%)
C-model	Dev	21.76 (3.88%)
	Test	21.81 (3.64%)
Oracle		15.58

Table 2: WER of the baseline system and after rescoring using prosodic models. Results are based on the first ASR system.

test data is smaller when using the *C-model* than *S-model*, which means that the prosodic model with co-training is more stable. In fact, we found that the optimal value of λ is 94 and 57 for the two folds using *S-model*, and is 99 and 110 for the *C-model*. These verify again that the prosodic scores contribute more in the combination with ASR likelihood scores when using the *C-model*, and are more robust across different tuning sets. Ananthakrishnan and Narayanan (2007) also used acoustic/lexical prosodic models to estimate a prosody score and reported 0.3% recognition error reduction on BU data when rescoring 100-best list (their baseline WER is 22.8%). Although there is some difference in experimental setup (data, classifier, features) between ours and theirs, our *S-model* showed comparable performance gain and the result of *C-model* is significantly better than theirs.

Next we test our n-best rescoring approach using a state-of-the-art SRI speech recognizer on BN data to verify if our approach can generalize to better ASR n-best lists. This is often the concern that improvements observed on a poor ASR system do not hold for better ASR systems. The rescoring results are shown in Table 3. We can see that the baseline performance of this recognizer is much better than that of the first ASR system (even though the recognition task is also harder). Our rescoring approach still yields performance gain even using this state-of-the-art system. The WER is reduced by 0.29% (2.07% relative). This error reduction is lower than that in the first ASR system. There are several possible reasons. First, the baseline ASR performance is higher, making further improvement hard; second, and more importantly, the prosody models do not match well to the test domain. We trained the prosody model using the BU data. Even though co-training is used to leverage unlabeled BN data to reduce data mismatch, it is still not as good as using labeled in-domain data for model training.

		WER (%)
1-best baseline		13.77
S-model	Dev	13.53 (1.78%)
	Test	13.55 (1.63%)
C-model	Dev	13.48 (2.16%)
	Test	13.49 (2.07%)
Oracle		9.23

Table 3: WER of the baseline system and after rescoring using prosodic models. Results are based on the second ASR system.

6.3 Analysis and Discussion

We also analyze what kinds of errors are reduced using our rescoring approach. Most of the error reduction came from substitution and insertion errors. Deletion error rate did not change much or sometimes even increased. For a better understanding of the improvement using the prosody model, we analyzed the pattern of corrections (the new hypothesis after rescoring is correct while the original 1-best is wrong) and errors. Table 4 shows some positive and negative examples from rescoring results using the first ASR system. In this table, each word is associated with some binary expressions inside a parenthesis, which stand for pitch-accent markers. Two bits are used for each syllable: the first one is for the acoustic-prosodic model and the second one is for the lexical-prosodic model. For both bits, 1 represents pitch-accent, and 0 indicates none. These hard decisions are obtained by setting a threshold of 0.5 for the posterior probabilities from the acoustic or lexical models. For example, when the acoustic classifier predicts a syllable as pitch-accented and the lexical one as not accented, '*10*' marker is assigned to the syllable. The number of such pairs of pitch-accent markers is the same as the number of syllables in a word. The bold words indicate correct words and italic means errors. As shown in the positive example of Table 4, we find that our prosodic model is effective at identifying an erroneous word when it is split into two words, resulting in different pitch-accent patterns. Language models are

Positive example	1-best :	most	*of*	*the*	massachusetts	
		(11)	***(10)***	***(00)***	(11 00 01 00)	
	rescored :	most	**other**		massachusetts	
		(11)	**(11 00)**		(11 00 01 00)	
Negative example	1-best :	robbery	and	**on**	**a**	theft
		(11 00 00)	(00)	**(10)**	**(00)**	(11)
	rescored :	robbery	and	*lot*	*of*	theft
		(11 00 00)	(00)	***(11)***	***(00)***	(11)

Table 4: Examples of rescoring results. Binary expressions inside the parenthesis below a word represent pitch-accent markers for the syllables in the word.

not good at correcting this kind of errors since both word sequences are plausible. Our model also introduces some errors, as shown in the negative example, which is mainly due to the inaccurate prosody model.

We conducted more prosody rescoring experiments in order to understand the model behavior. These analyses are based on the n-best list from the first ASR system for the entire test set. In the first experiment, among the 100 hypotheses in n-best list, we gave a prosody score of 0 to the 100^{th} hypothesis, and used automatically obtained prosodic scores for the other hypotheses. A zero prosody score means the perfect agreement given acoustic and lexical cues. The original scores from the recognizer were combined with the prosodic scores for rescoring. This was to verify that the range of the weighting factor λ estimated on the development data (using the original, not the modified prosody scores for all candidates) was reasonable to choose proper hypothesis among all the candidates. We noticed that 27% of the times the last hypothesis on the list was selected as the best hypothesis. This hypothesis has the highest prosodic scores, but lowest ASR score. This result showed that if the prosodic models were accurate enough, the correct candidate could be chosen using our rescoring framework.

In the second experiment, we put the reference text together with the other candidates. We use the same ASR scores for all candidates, and generated prosodic scores using our prosody model. This was to test that our model could pick up correct candidate using only the prosodic score. We found that for 26% of the utterances, the reference transcript was chosen as the best one. This was significantly better than random selection (i.e., 1/100), suggesting the benefit of the prosody model; however, this percentage is not very high, implying the limitation of prosodic information for ASR or the current imperfect prosodic models.

In the third experiment, we replaced the 100^{th} candidate with the reference transcript and kept its ASR score. When using our prosody rescoring approach, we obtained a relative error rate reduction of 6.27%. This demonstrates again that our rescoring method works well – if the correct hypothesis is on the list, even though with a low ASR score, using prosodic information can help identify the correct candidate.

Overall the performance improvement we obtained from rescoring by incorporating prosodic information is very promising. Our evaluation using two different ASR systems shows that the improvement holds even when we use a state-of-the-art recognizer and the training data for the prosody model does not come from the same corpus. We believe the consistent improvements we observed for different conditions show that this is a direction worthy of further investigation.

7 Conclusion

In this paper, we attempt to integrate prosodic information for ASR using an n-best rescoring scheme. This approach decouples the prosodic model from the main ASR system, thus the prosodic model can be built independently. The prosodic scores that we use for n-best rescoring are based on the matching of pitch-accent patterns by acoustic and lexical features. Our rescoring method achieved a WER reduction of 3.64% and 2.07% relatively using two different ASR systems. The fact that the gain holds across different baseline systems (including a state-of-the-art speech recognizer) suggests the possibility that prosody can be used to improve speech recognition performance.

As suggested by our experiments, better prosodic models can result in more WER reduction. The performance of our prosodic model was improved with co-training, but there are still problems, such as the imbalance of the two classifiers' prediction, as well as for the two events. In order to address these problems, we plan to improve the labeling and selection method in the co-training algorithm, and also explore other training algorithms to reduce domain mismatch. Furthermore, we are also interested in evaluating our approach on the spontaneous speech domain, which is quite different from the data we used in this study.

In this study, we used n-best rather than lattice rescoring. Since the prosodic features we use include cross-word contextual information, it is not straightforward to apply it directly to lattices. In our future work, we will develop models with only within-word context, and thus allowing us to explore lattice rescoring, which we expect will yield more performance gain.

References

Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan. 2007. Improved speech recognition using acoustic and lexical correlated of pitch accent in a n-best rescoring framework. *Proc. of ICASSP*, pages 65–68.

Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan. 2008. Automatic prosodic event detection using acoustic, lexical and syntactic evidence. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):216–228.

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes. *Proc. of NAACL-HLT*, pages 308–316.

Stefan Benus, Agustín Gravano, and Julia Hirschberg. 2007. Prosody, emotions, and whatever. *Proc. of Interspeech*, pages 2629–2632.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. *Proc. of the Workshop on Computational Learning Theory*, pages 92–100.

Ken Chen and Mark Hasegawa-Johnson. 2006. Prosody dependent speech recognition on radio news corpus of American English. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):232– 245.

Najim Dehak, Pierre Dumouchel, and Patrick Kenny. 2007. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2095–2103.

Esther Grabe, Greg Kochanski, and John Coleman. 2003. Quantitative modelling of intonational variation. *Proc. of SASRTLM*, pages 45–57.

Je Hun Jeon and Yang Liu. 2009. Automatic prosodic events detection suing syllable-based acoustic and syntactic features. *Proc. of ICASSP*, pages 4565–4568.

Je Hun Jeon and Yang Liu. 2010. Syllable-level prominence detection with acoustic evidence. *Proc. of Interspeech*, pages 1772–1775.

Ozlem Kalinli and Shrikanth Narayanan. 2009. Continuous speech recognition using attention shift decoding with soft decision. *Proc. of Interspeech*, pages 1927–1930.

Diane J. Litman, Julia B. Hirschberg, and Marc Swerts. 2000. Predicting automatic speech recognition performance using prosodic cues. *Proc. of NAACL*, pages 218–225.

Mari Ostendorf, Patti Price, and Stefanie Shattuck-Hufnagel. 1995. The Boston University radio news corpus. *Linguistic Data Consortium.*

Mari Ostendorf, Izhak Shafran, and Rebecca Bates. 2003. Prosody models for conversational speech recognition. *Proc. of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, pages 147–154.

Andrew Rosenberg and Julia Hirschberg. 2006. Story segmentation of broadcast news in English, Mandarin and Arabic. *Proc. of HLT-NAACL*, pages 125–128.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154.

Elizabeth Shriberg, Luciana Ferrer, Sachin S. Kajarekar, Anand Venkataraman, and Andreas Stolcke. 2005. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, 46(3-4):455–472.

Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth S. Narayanan. 2008. Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):797–811.

Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth Narayanan. 2009. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech and Language*, 23(4):407–422.

Andreas Stolcke, Barry Chen, Horacio Franco, Venkata Ramana Rao Gadde, Martin Graciarena, Mei-Yuh Hwang, Katrin Kirchhoff, Arindam Mandal, Nelson Morgan, Xin Lin, Tim Ng, Mari Ostendorf, Kemal Sönmez, Anand Venkataraman, Dimitra Vergyri, Wen Wang, Jing Zheng, and Qifeng Zhu. 2006. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1729–1744. Special Issue on Progress in Rich Transcription.

Gyorgy Szaszak and Klara Vicsi. 2007. Speech recognition supported by prosodic information for fixed stress languages. *Proc. of TSD Conference*, pages 262–269.

Dimitra Vergyri, Andreas Stolcke, Venkata R. R. Gadde, Luciana Ferrer, and Elizabeth Shriberg. 2003. Prosodic knowledge sources for automatic speech recognition. *Proc. of ICASSP*, pages 208–211.

Colin W. Wightman and Mari Ostendorf. 1994. Automatic labeling of prosodic patterns. *IEEE Transaction on Speech and Auido Processing*, 2(4):469–481.

Jing Zheng, Ozgur Cetin, Mei-Yuh Hwang, Xin Lei, Andreas Stolcke, and Nelson Morgan. 2007. Combining discriminative feature, transform, and model training for large vocabulary speech recognition. *Proc. of ICASSP*, pages 633–636.