IMPROVED SPEECH RECOGNITION USING ACOUSTIC AND LEXICAL CORRELATES OF PITCH ACCENT IN A N-BEST RESCORING FRAMEWORK

Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory
Department of Electrical Engineering
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089

ananthak@usc.edu, shri@sipi.usc.edu

ABSTRACT

Most statistical speech recognition systems make use of segment-level features, derived mainly from spectral envelope characteristics of the signal, but ignore supra-segmental cues that carry additional information likely to be useful for speech recognition. These cues, which constitute the prosody of the utterance and occur at the syllable, word and utterance level, are closely related to the lexical and syntactic organization of the utterance. In this paper, we explore the use of acoustic and lexical correlates of a subset of these cues in order to improve recognition performance on a read-speech corpus, using word error rate (WER) as the metric. Using the features and methods described in this paper, we were able to obtain a relative WER improvement of 1.3% over a baseline ASR system on the Boston University Radio News Corpus.

Index Terms— speech recognition, prosody, re-ranking N-best lists

1. INTRODUCTION

Most modern automatic speech recognition (ASR) systems consist of two components: a) an acoustic model, which provides the likelihood of acoustic (spectral) features given the sequence of words, and b) a language model, which establishes constraints on hypothesized word sequences. ASR acoustic features are usually extracted over window lengths of a few tens of milliseconds, and while they are well-suited for capturing the variation of spectral content over time, they miss useful information contained in the prosody of the utterance, which manifests itself at higher linguistic (syllable, word, and utterance) levels. Likewise, language models typically used in ASR do not exploit the relationship between lexical items and the prosodic structure of the utterance.

Previous work on automatic annotation of prosodic events (such as pitch accents and phrase boundaries) in speech [1, 2] has demonstrated that there is a close relationship between said events and the lexico-syntactic structure of the utterance. In [2], we showed that pitch accents are strongly correlated with syllable tokens that occur mostly in content words. Moreover, canonical stress patterns annotated in many pronunciation dictionaries are good indicators of pitch accents in speech. These dependencies can be used to augment the standard ASR model to improve recognition performance.

The task of integrating prosody within an ASR framework has been previously dealt with in [3], [4] and [5]. In general, this is a difficult problem, since prosodic events occur over larger, ill-defined time-scales, giving rise to problems of asynchronicity. One way is to incorporate prosodic features as another stream at the segment level [5]; this has the advantage that spectral and prosodic features are jointly modeled. However, with this framework, we cannot capture phenomena beyond the segment level. Alternatively, models for prosody can be built independent of the ASR acoustic and language models. This has two advantages: a) the models can be built at arbitrary linguistic levels and combined with the ASR hypotheses as a post-processing step (lattice or N-best list rescoring) and b) no modification of the conventional ASR is necessary to include prosodic information. The downside of this approach is that correlations between the spectral and prosodic features will not be captured by the model. Yet another consideration is whether the relation between prosodic and lexical elements is modeled directly, or through an intermediate symbolic transcription (such as ToBI [6] or its subsets).

In this paper, we investigate the use of prosodic events in the form of pitch accents to improve speech recognition over a baseline ASR system. We use a simple binary intermediate symbolic representation of pitch accents in the form of binary "yes"-"no" tags, derived from a ToBI-style transcription of prosody. We adopt the N-best rescoring approach, assuming that the acoustic-prosodic features are conditionally independent of the spectral features given the word sequence and pitch accent events. The rest of this paper is organized as fol-

lows: Section 2 describes our data corpus and baseline ASR setup. Section 3 presents our prosody model and the N-best list re-ranking scheme. Section 4 describes the results of our re-ranking experiments. Finally, Section 5 includes a brief discussion of the work presented in this paper and outlines future directions for research.

2. DATA CORPUS AND BASELINE ASR

The Boston University Radio News Corpus (BU-RNC) [7] consists of about 3 hours of read speech with 6 speakers (3 female, 3 male). We use this corpus because it contains prosodic annotations in the form of ToBI-style labels for pitch accents, phrase boundaries and lexical break indices. After eliminating news story repetitions by the same speaker, the remaining data (about 2h 40m worth of speech) was split into 10 random training and evaluation partitions which were approximately equal in size (14K words vs. 13K words). The evaluation partitions were further divided into held-out development (4.2K words) and test sets (8.7K words). We developed a baseline ASR for this corpus as described below. No prosodic information was used in the design of the baseline system.

2.1. Baseline ASR

We used the University of Colorado SONIC continuous speech recognizer [8] to develop the baseline ASR. We adapted context-dependent triphone acoustic models from the Wall Street Journal (WSJ) task with data from the training partitions of the BU-RNC using the tree-based MAPLR algorithm supported by SONIC. The adapted acoustic models were gender-specific but otherwise speaker independent. We used PMVDR features derived from the acoustic signal to train these models. A standard back-off trigram language model with Kneser-Ney smoothing was trained with a mixture of text from the WSJ, HUB-4 and BU datasets (totaling over 4.7 million words). The language model vocabulary was slightly over 28.5K words; the test set vocabulary was approximately 2.3K words. The out-of-vocabulary (OOV) rate on the test set was 2.0%. The baseline ASR was used to generate 1-best hypotheses for the evaluation utterances.

2.2. N-best list generation

In addition to the baseline 1-best hypothesis, we also generated a N-best list containing the top-N hypotheses for each evaluation utterance. After analysis of oracle WER drop-off rates for different values of N (see Section 4), we chose $N = 100$ as a good compromise between processing time and potential loss of accuracy.

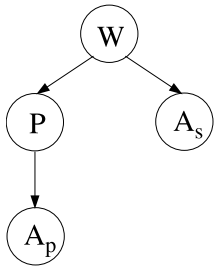

Fig. 1. Directed graph illustrating conditional independence assumptions between spectral features and prosodic events, given the sequence of words.

3. PROSODY MODEL

We attempt to exploit the relationship between pitch accents and lexical items to improve ASR performance. We augment the standard ASR equation to include prosodic information as follows.

$$(\mathbf{W}^*, \mathbf{P}^*) = \arg\max_{\mathbf{W},\mathbf{P}} p(\mathbf{W}, \mathbf{P} | \mathbf{A_s}, \mathbf{A_p}) \qquad (1)$$

This is similar to the models presented in [3] and [5]. However, our subsequent decomposition and architecture of this model is different in the following respects.

- We do not modify the original ASR acoustic or language models. Prosody is used as an external knowledge source that can be used to refine ASR hypotheses in a N-best or lattice rescoring framework.
- Our prosody model is constructed at the sub-word (linguistic syllable) level, since pitch accents are carried by syllables. This makes it better suited for disambiguating words which have the same phonetic pronunciation, but carry pitch accents on different syllables.

Based on conditional independence assumptions encoded by Figure 1, Eq. 1 can be rewritten as follows.

$$\begin{aligned}
(\mathbf{W}^*, \mathbf{P}^*) &= \arg\max_{\mathbf{W},\mathbf{P}} p(\mathbf{W}, \mathbf{P}, \mathbf{A_s}, \mathbf{A_p}) \\
&= \arg\max_{\mathbf{W},\mathbf{P}} p(\mathbf{W}) p(\mathbf{A_s}|\mathbf{W}) p(\mathbf{P}|\mathbf{W}) p(\mathbf{A_p}|\mathbf{P}) \\
&= \arg\max_{\mathbf{W},\mathbf{P}} \underbrace{p(\mathbf{W}) p(\mathbf{A_s}|\mathbf{W})}_{\text{ASR score}} \cdot \underbrace{\frac{p(\mathbf{P}|\mathbf{A_p})}{p(\mathbf{P})} p(\mathbf{P}|\mathbf{W})}_{\text{prosody score}}
\end{aligned}$$

where $\mathbf{W}$, $\mathbf{A_s}$, $\mathbf{P}$ and $\mathbf{A_p}$ stand for the word, acoustic (spectral) feature, pitch accent label, and acoustic-prosodic feature sequences, respectively. The prosody model was built at the syllable level. Syllable-level transcriptions of the training data and N-best hypotheses were obtained by running

a rule-based syllabifier [9] on the text. The pitch accent label sequence was obtained by binarizing all ToBI-style pitch accents to "yes"-"no" categories. Acoustic-prosodic features were extracted from the speech signal from automatic syllable-level forced-alignment of the training text or the N-best hypotheses, depending on whether we were training or evaluating the system. The model shown above has three subcomponents, which are described as follows.

3.1. Acoustic-prosodic model

This refers to the model $p(\mathbf{P}|\mathbf{A_p})$, which provides the posterior probability of pitch accent labels given the acoustic-prosodic evidence. Based on previous work in prosody labeling, the acoustic-prosodic features that make up $\mathbf{A_p}$ include

1. F0: F0-range features (max-min, max-avg, avg-min), difference in mean F0 between current, previous and next syllable

2. Energy: within-syllable energy range features (max-min, avg-min)

3. Timing: syllable nucleus duration

These features were normalized to minimize effects of speaker- or nucleus-specific variation. The model is trained as a feedforward neural network (MLP) with 8 input nodes, 25 hidden nodes and 2 output nodes with softmax activation, with outputs interpreted as posterior probabilities.

3.2. De-lexicalized prosody sequence model

The term $p(\mathbf{P})$ establishes constraints on the sequence of pitch accent events $\mathbf{P}$. Since $\mathbf{P}$ has a binary vocabulary, it was robustly estimated from small amounts of training data. We modeled this component as a 4-gram back-off language model with pitch accent labels obtained from the training data.

3.3. Lexical prosody sequence model

The lexical prosody sequence model $p(\mathbf{P}|\mathbf{W})$ establishes constraints on the sequence of prosody labels $\mathbf{P}$ given the word sequence $\mathbf{W}$. Since we built prosody models at the syllable level, we first decomposed the sequence of words into the corresponding sequence of syllables $\mathbf{S}$ using the syllabifier. Concurrently, we obtained a canonical stress label for each syllable from the widely available CMU pronunciation dictionary. We have previously shown [2] that these canonical stress labels exhibit high correlation with pitch accents. This provided us with another stream of features $\mathbf{L}$. The lexical prosody sequence model then becomes $p(\mathbf{P}|\mathbf{W}) = p(\mathbf{P}|\mathbf{S},\mathbf{L})$. We approximated this as a language model with multiple factors, as shown below for a bigram structure. In practice, we used a trigram structure for this model.

$$p(\mathbf{P}|\mathbf{S},\mathbf{L}) = p(p_1|s_1,l_1) \cdot \prod_{i=2}^{n} p(p_i|s_i,l_i,p_{i-1},s_{i-1},l_{i-1})$$

Each of the above prosody models was trained using slightly over 22,800 syllable samples. The sequence models were trained with explicit word boundary tags. The parameter-rich lexical prosody sequence model presented a sparsity issue, since we only had a small amount of annotated data with which to train this model. In order to obtain smoothed probabilities for this model, we implemented it as a factored back-off LM [10] with a fixed back-off path. Due to their considerable vocabulary, the greatest effect of sparsity was on account of the syllable tokens s_i; at any given level in the history, we dropped these first in our back-off structure. We used the SRILM toolkit [11] to train this model.

4. EXPERIMENTS AND RESULTS

The entire data corpus was divided into training, held-out development and test sets, and a baseline ASR was built according to the description given in Section 2. The average 10-fold cross-validated baseline WER on the test set was determined to be 22.8% (22.7% on the development set). We also generated N-best lists for each test utterance for use in our re-ranking experiments. The upper curve in Figure 2 shows the variation in oracle WER as a function of N for the baseline N-best lists. It is clear that the greatest improvement in oracle WER occurs at lower values of N. Based on this empirical observation, we set $N = 100$. The average oracle WER for these 100-best lists across 10 cross-validation test sets was 19.8%, and the average anti-oracle WER was 27.0%. This represented a 3.0% margin for improvement, and a 4.2% margin for degradation.

The prosody model was evaluated on each hypothesis of the N-best lists, which were re-ranked based on a weighted combination of the ASR score (generated from the acoustic and language models) and the score assigned by the prosody model. The weight of the prosody model was optimized on the held-out development data and was then applied to the test data. After re-ranking, the average 1-best WER on the development set reduced by 0.4% to 22.3%, while the average 1-best WER on the test set was 22.5%, corresponding to a 0.3% absolute (1.3% relative) reduction in WER. Table 1 summarizes these results, while the lower curve in Figure 2 illustrates the variation in oracle WER as a function of N for the re-ranked N-best lists.

In order to determine whether the improvement was statistically significant, we used the Wilcoxon signed rank test, which is a non-parametric method for comparing matched pairs and reporting whether their differences originate from a zero-median distribution. According to this test, the difference between baseline and re-ranked WER was significant at

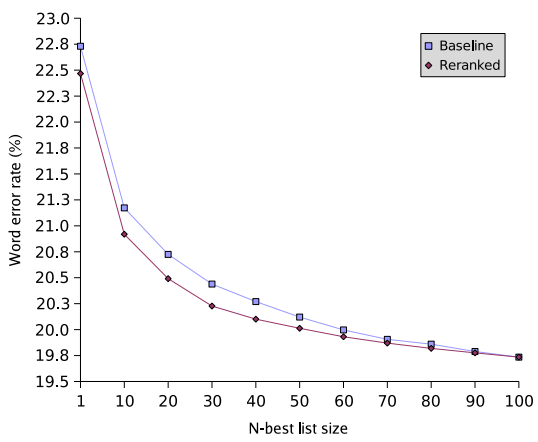

Fig. 2. Oracle WER on the test set for baseline and re-ranked N-best lists as a function of N. Upper curve is the baseline system, lower curve is the re-ranked system.

the $p \leq 0.002$ level. This was corroborated by the fact that there was a modest but consistent improvement in WER in each of the 10 cross-validation test sets.

5. DISCUSSION

In this paper, we presented a N-best re-ranking scheme using a prosody model that was decoupled from the main ASR system. The re-ranking method achieved a modest but significant reduction in WER of 1.3% (relative) compared to the baseline recognition system. We directly modeled the relationship between binary pitch accent labels, acoustic-prosodic features, and lexical items (syllable tokens) without the need for other sources of information, such as part-of-speech. The structure of our prosody model makes it possible to integrate prosodic information within ASR-generated lattices or word meshes without the need to produce N-best lists and without the concomitant loss of information.

We obtained results comparable to [5], where the authors presented a baseline recognizer with a WER of 24.8%, which improved to 21.7% with their best performing system. Our baseline WER was 2% lower, and our testing conditions were more stiff in the following respects: (a) we did not permit story repetitions by the same speaker to co-exist in training and test data, (b) we used only 50% of the remaining data for training ([5] used 90% of the data for training), and (c) we used only pitch accent in our prosody model.

One limitation of this method is that it requires training data to be hand-annotated with the prosodic events of interest. In order to alleviate this problem, we are investigating unsupervised prosodic event detection techniques based on clustering algorithms [12], which we hope to use to improve ASR performance without the need for hand-labeled data.

Table 1. ASR performance

System	Dev. WER	Test WER	Significance
Baseline	22.7%	22.8%	
Reranked	22.3%	22.5%	$p \leq 0.002$

6. REFERENCES

[1] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2004, pp. 509–512.

[2] S. Ananthakrishnan and S. Narayanan, "Automatic prosody labeling using acoustic, lexical and syntactic evidence," *submitted to the IEEE Transactions on Speech and Audio Processing*, 2006.

[3] M. Ostendorf, I. Shafran, and R. Bates, "Prosody models for conversational speech recognition," in *Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, 2003.

[4] C. Wang and S. Seneff, "Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the jupiter domain," in *Proceedings of the 7th European Conference on Speech Communication and Technology*, 2001.

[5] M. Hasegawa-Johnson, J. Cole, C. Shih, K. Chen, A. Cohen, S. Chavarria, H. Kim, T. Yoon, S. Borys, and J.-Y. Choi, "Speech recognition models of the interdependence among syntax, prosody and segmental acoustics," in *Proceedings of HLT/NAACL*, 2004.

[6] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard scheme for labeling prosody," in *Proceedings of the International Conference on Spoken Language Processing*, 1992, pp. 867–869.

[7] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," 1995.

[8] B. Pellom, "Sonic: The University of Colorado continuous speech recognizer," University of Colorado, Tech. Rep. TR-CSLR-2001-01, March 2001.

[9] D. Kahn, "Syllable-based generalizations in English phonology," Ph.D. dissertation, University of Massachusetts, 1976.

[10] J. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proceedings of HLT/NAACL*, 2003.

[11] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of the International Conference of Spoken Language Processing*, 2002.

[12] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling," in *Proceedings of the International Conference on Spoken Language Processing*, 2006, pp. 297–300.