Prosody-enriched lattices for improved syllable recognition

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

Automatic recognition of syllables is useful for many spoken language applications such as speech recognition and spoken document retrieval. Short-term spectral properties (such as mel-frequency cepstral coefficients, or MFCCs) are usually the features of choice for such systems, which typically ignore supra-segmental (prosodic) cues that manifest themselves at the syllable, word and utterance level. Previous work has shown that categorical representations of prosody correlate well with lexical entities. In this paper, we attempt to exploit this relationship by enriching syllable-level lattices, generated by a standard speech recognizer, with categorical prosodic events for improved syllable recognition performance. With the enriched lattices, we obtain a 2% relative improvement in syllable error rate over the baseline system on a read speech task (the Boston University Radio News Corpus).

Index Terms: syllable recognition, prosody, pitch accent, enriched lattices

1. Introduction

Although most automatic speech recognition (ASR) systems produce word hypotheses as their output, syllable and related sub-word unit recognition is important for several spoken language applications. Speech recognizers that produce word hypotheses perform poorly on tasks such as name recognition, since names and other proper nouns account for most of the out-of-vocabulary (OOV) terms in such systems. Phone- and syllable-recognizer based approaches are therefore preferred. On a spoken name recognition task, Sethy et al. [1] report a significant performance improvement over a phone-based system by using hybrid syllable and phone models. Spoken document retrieval (SDR) is another domain where sub-word units such as syllables are useful. Many SDR systems use syllables [2] and phone n-grams as indexing features in situations where words are not suitable (for instance, where the domain is unknown or too varied, resulting in a high OOV rate).

A standard speech recognition system is comprised of a) an acoustic model that provides the likelihood of segment-level spectral features (MFCCs) given the lexical items, which in this case are syllables and b) a language model that establishes constraints on the sequence of lexical items by assigning high probability to sequences that are more likely to be seen, and vice-versa. While this architecture works reasonably well for the purposes of transcribing speech, it completely ignores the higher level prosodic structure of the utterance, which contains information independent of the segment-level features. Prosody encodes different types of syntactic constructs through the introduction of accents, pauses and hesitations (boundaries). Various speech acts (questions, exclamations, declaratives) are also enriched by these supra-segmental effects. However, traditional speech recognition systems by and large do not exploit this additional source of information.

Two of the biggest stumbling blocks to the incorporation of prosody in speech recognition are: a) the asynchronous nature of key acoustic-prosodic features (F0, energy), which implies that these features are not time-aligned to the lexical entities of interest and b) the ill-defined relationship between raw acoustic-prosodic features and the lexical items that we wish to recognize. While the first issue remains a problem, the second has been partially addressed through the development of categorical representations of prosody that have a linguistic basis. It has been shown in Chen et al. [3] and Ananthakrishnan et al. [4] that categorical representations of prosody, such as those based on the Tones and Break Indices (ToBI) standard, share a close relationship with the lexical items (syllable tokens) of interest. For instance, pitch accents impart emphasis or prominence to specific syllables and usually co-occur with syllables that are part of content words, whereas syllables that commonly occur in function words are seldom associated with these events.

Previous work that exploits such dependencies between categorical prosodic events and lexical items to improve word recognition performance is described in Ostendorf et al. [5], Wang et al. [6], Hasegawa-Johnson et al. [7] and Ananthakrishnan et al. [8]. The authors of [7] described a system that augments segment-level acoustic models with acoustic-prosodic features and an enriched language model that incorporates categorical prosody labels derived from ToBI. In [8], we presented a prosody-based N-best list rescoring technique that exploits the correlation between pitch accent events and words to improve word recognition performance. This system does not require any modification of the existing ASR models.

In this paper, we explore the possibility of improving syllable recognition performance by enriching syllable lattices, generated by a standard ASR, with categorical prosodic events in order to exploit their relationship with syllable tokens. The advantage of using lattices over N-best lists is that conversion from the former to the latter is a lossy process, and there is a possibility that even if the correct hypothesis is present in the lattice, it may be pruned out of the N-best list unless N is very large. Pitch accent events are naturally suited to this task, since syllables are traditionally considered the smallest lexical units that carry these events. The pitch accent events used in our

approach are a simplified version of ToBI-style prosodic transcriptions, and are obtained by mapping the different tone tier accents to simple binary presence vs. absence categories.

The remainder of this paper is organized as follows. In Section 2, we provide a brief description of the recognition task and data corpus on which we carry out experiments. We also describe the development of the baseline syllable recognition system that does not use prosodic information. Section 3 contains a discussion of the prosody model that is used to augment the standard ASR-generated lattices. Section 4 describes the process by which syllable-level lattices are enriched with information from the prosody models. In Section 5, we discuss our experimental setup and present syllable recognition results for the prosody-enriched system. Finally, Section 6 contains a brief discussion of the work presented in this paper, and outlines future directions for research in this area.

2. Data corpus and baseline recognizer

We make use of the Boston University Radio News Corpus (BU-RNC) [9] for our syllable recognition experiments. This is a broadcast news style, read-speech corpus with speech data from 6 speakers (3 male and 3 female). The principal reason for working with this corpus is that, in addition to about 3 hours of speech data, most of the news stories in this corpus have been manually annotated for ToBI-style categorical prosodic events, including pitch accents and prosodic phrase boundaries. This makes it possible to train our prosody models in a supervised fashion. Besides this information, the corpus also provides F0 tracks and part-of-speech (POS) tags; these features have made it a standard corpus for work on prosody-enabled spoken language systems. From a lexical point of view, the BU-RNC contains approximately 48,800 syllables.

2.1. Baseline syllable recognizer

We first divided the BU-RNC into training and test partitions that were approximately equal in size. After eliminating story repetitions between the training and test sets, the training partition was comprised of approximately 22,800 syllables, while the test partition contained about 21,300 syllables. We used the University of Colorado SONIC [10] continuous speech recognizer to develop the baseline syllable recognizer. In this paper, our approach is to treat syllables simply as phone strings. We therefore developed phone-level acoustic models for the syllable recognition task by adapting acoustic models from the Wall Street Journal (WSJ) task with data from the BU-RNC. An alternative approach would be to develop the acoustic models at the syllable level (depending on the availability of training data), or to work with a hybrid set comprised of both syllable and phone models. Adaptation was carried out with the tree-based MAPLR algorithm supported by SONIC. The baseline system used segment-level MFCCs as acoustic features. Our acoustic models were speaker independent except for the fact that we developed separate acoustic models for male and female speakers.

In order to build the syllable-level language model, we syllabified word-level transcripts in the BU-RNC using a deterministic syllabification algorithm [11] based on the phonological rules of English. We then trained a back-off trigram language model from these syllabified transcripts using the SRILM [12] toolkit.

The test utterances were then decoded, using the acoustic and language models described above, in order to generate the baseline lattices. In addition to syllable tokens, acoustic and language model scores, these lattices also contain syllable and constituent phone-level time alignments. Viterbi decoding of these lattices produced a 1-best syllable error rate of 30.6%.

3. Prosody-enabled syllable recognizer

The prosody-enabled syllable recognition architecture is developed in the same manner as [8] and is similar to the formulations presented in [5, 7]. In this case, the decomposition and simplification is somewhat different from our previous work. We augment the standard ASR maximum a-posteriori equation to include categorical prosody labels (pitch accents) and acoustic-prosodic features as shown in Eq. 1.

$$
\begin{aligned}
(\mathbf{S}^*, \mathbf{P}^*) &= \arg\max_{\mathbf{S}, \mathbf{P}} p(\mathbf{S}, \mathbf{P} | \mathbf{A_s}, \mathbf{A_P}) \\
&= \arg\max_{\mathbf{S}, \mathbf{P}} p(\mathbf{S}, \mathbf{P}, \mathbf{A_s}, \mathbf{A_P}) \quad (1)
\end{aligned}
$$

where $\mathbf{S}$ denotes the sequence of syllables, $\mathbf{P}$ the sequence of prosody labels, $\mathbf{A_s}$ the segment-level spectral features (MFCCs) and $\mathbf{A_P}$ the acoustic-prosodic features. As the joint probability density function is too complex to model with sparse data, we invoke conditional independence assumptions that make the problem more tractable. Specifically, we assume that

- the prosody labels $\mathbf{P}$ are conditionally independent of the segment-level features $\mathbf{A_s}$ given the syllable sequence $\mathbf{S}$.

- the acoustic-prosodic features $\mathbf{A_P}$ are conditionally independent of the syllable sequence $\mathbf{S}$ and segment-level features $\mathbf{A_s}$ given the prosody labels $\mathbf{P}$.

Upon making these assumptions, Eq. 1 simplifies to

$$
\begin{aligned}
(\mathbf{S}^*, \mathbf{P}^*) &= \arg\max_{\mathbf{S}, \mathbf{P}} p(\mathbf{S}) p(\mathbf{A_s}|\mathbf{S}) p(\mathbf{P}|\mathbf{S}) p(\mathbf{A_P}|\mathbf{P}) \\
&= \arg\max_{\mathbf{S}, \mathbf{P}} \underbrace{p(\mathbf{A_s}|\mathbf{S})}_{\text{AM score}} \cdot \underbrace{p(\mathbf{P}|\mathbf{A_P}) p(\mathbf{S}|\mathbf{P})}_{\text{prosody score}} \quad (2)
\end{aligned}
$$

In [8], we retain both the acoustic and language model scores provided by the ASR. However, in the above formulation, we retain only the acoustic model score and replace the language model by a prosody-enabled language model $p(\mathbf{S}|\mathbf{P})$. The prosodic score in Eq. 2 has two components, which we describe below.

3.1. Acoustic-prosodic model

The acoustic-prosodic model $p(\mathbf{P}|\mathbf{A_P})$ provides the posterior probability of a prosodic event $\mathbf{P}$ given the acoustic-prosodic features $\mathbf{A_P}$. In this case, prosodic events refer to binary presence vs. absence labels for pitch accent. The acoustic-prosodic features are derived from the speech signal; they are based on previous work on automatic detection of pitch accents [4] and include

1. F0: F0-range features (max-min, max-avg, avg-min), difference between within-syllable mean F0 and utterance average F0

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

3. Timing: syllable nucleus duration

The choice of range and difference features for F0 and energy serves to normalize the acoustic-prosodic features against
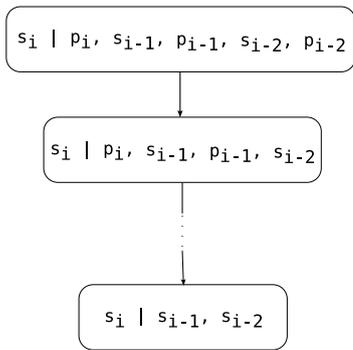
Figure 1: Back-off structure: factored prosodic language model

speaker- or gender-specific variation while retaining class-specific (presence vs. absence of pitch accent) variation. The syllable nucleus duration is normalized as described in [4] to eliminate the intrinsic duration variation due to vowel-type and again retain only class-specific variation. The features are extracted using the syllable-level time-alignments from the ASR-generated lattices.

The acoustic-prosodic model functions as a binary classifier and is trained as a feedforward neural network (MLP) [13] with 8 input nodes, 25 hidden nodes and 2 output nodes. Softmax activation is used for the output nodes; this enables us to treat the output of the neural network as class posterior probabilities.

3.2. Prosodic language model

The prosodic language model $p(\mathbf{S}|\mathbf{P})$ establishes constraints on the syllable sequence $\mathbf{S}$ given the sequence of prosody labels $\mathbf{P}$. In this model, the probability of the current syllable token depends not only on the history of syllable tokens (a regular n-gram), but also on the current and past prosody labels as shown in Eq. 3 for trigram context.

$$p(\mathbf{S}|\mathbf{P}) = \prod p(s_i|p_i, s_{i-1}, p_{i-1}, s_{i-2}, p_{i-2}) \qquad (3)$$

In order to alleviate sparsity issues due to the increased number of conditioning factors and a relatively small training set, we implemented this component as a factored back-off language model (trigram context), with the syllable tokens and prosody labels as the conditioning factors. We chose a fixed back-off path for this model, beginning with full syllable token and prosody label history and dropping the conditioning prosody labels at each step as shown in Figure 1. The final back-off node contains no prosodic conditioning factors and correspond to a regular syllable n-gram structure. This model was trained with the factored LM tools that are part of the SRILM package.

4. Enriched syllable lattices

The lattices generated by the baseline recognizer encode many possible syllable sequence hypotheses, with the syllable tokens forming the arcs of the lattices. Associated with each arc is an acoustic score and a language model score from the baseline system. Also associated with each syllable token are syllable and phone-level time alignments that mark the start and end times of each hypothesized unit.

In order to enrich the baseline lattices with the prosody models described in the previous section, we use the lattice time-alignments to extract acoustic-prosodic features for each

Table 1: ASR performance

System	Test SER	Significance
Baseline	30.6%	
Enriched	30.0%	$p \le 0.001$

arc. The acoustic-prosodic model uses these features to compute the posterior probability of a pitch accent for each arc. Each arc in the baseline lattice is then replaced by multiple arcs that carry compound tokens constructed from the syllable token and the prosody label. The acoustic model score associated with each arc is adjusted according to the posterior probability returned by the acoustic-prosodic model. For instance, if the baseline lattice contains an arc that carries the syllable S, with acoustic model likelihood A, the enriched lattice contains in its place two arcs that carry compound tokens *S:0* with compound score $A(1 - Q)$ and *S:1* with compound score AQ, where Q is the posterior probability of the syllable S carrying a pitch accent. This operation is easily accomplished, since the baseline lattices carry syllable tokens on the arcs. Figure 2(a) represents a segment of a sample baseline lattice. The prosody-enriched version of the same segment is shown in Figure 2(b). Although we work only with binary pitch accent labels in this paper, the above method easily generalizes to arbitrary prosodic event categories.

At this point, the lattice incorporates scores from the baseline acoustic model and from the acoustic-prosodic model. The relationship between pitch accent labels and syllable tokens is then established by simply replacing the baseline language model weights by the appropriate prosodic language model weights. Viterbi decoding of these enriched lattices implements the augmented syllable recognizer of Eq. 2.

5. Experimental results

We first split the BU-RNC data into training and test partitions, and developed the baseline syllable recognizer as described in Section 2. The syllable error rate (SER) of this system based on Viterbi decoding of unenriched lattices was found to be 30.6%.

The baseline syllable lattices were then enriched with categorical prosody labels according to the description in Section 4. Since we work with binary prosody labels (presence vs. absence of pitch accent), each arc in the baseline lattices was replaced by two compound arcs with appropriately modified acoustic weights. Prosodic language model weights were then applied using the SRILM toolkit to obtain the prosody enriched lattices. Viterbi decoding of the enriched lattices resulted in a 1-best error rate of 30.0%, which represented a 0.6% absolute improvement (2% relative improvement) over the baseline.

We used the NIST matched-pairs sentence segment word error (MAPSSWE) [14] test in order to determine whether the improvement due to enriched lattices was statistically significant. This is an effective and commonly used test for comparing two speech recognition systems. According to this test, the performance improvement obtained with the prosody-enriched lattices was significant at the $p \le 0.001$ level. These results are summarized in Table 1.

6. Discussion and future directions

In this paper, we described our approach to generate syllable-level lattices enriched with categorical prosodic information for improved recognition performance. With the enriched lattices,
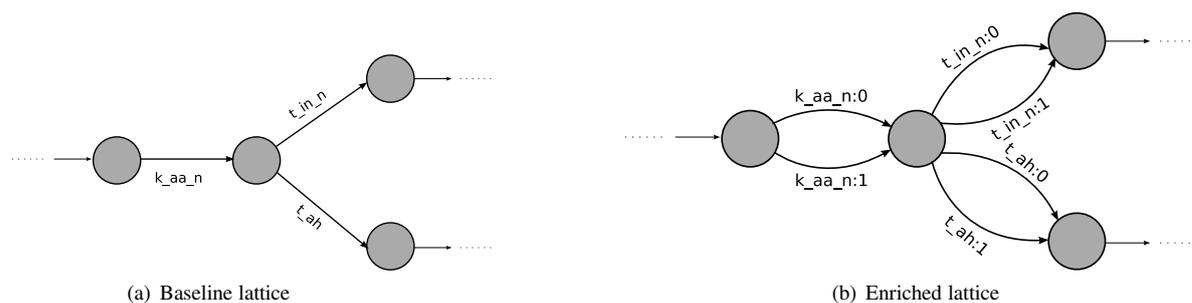
Figure 2: Sample baseline and prosody-enriched lattices. Binary prosody labels are used in this example.

we were able to obtain a 2% relative reduction in SER over the baseline; this improvement was determined to be statistically significant. The prosody models used to enrich the lattices were more or less decoupled from the baseline system, except for the fact that they depended on ASR-generated time-alignments for extracting acoustic-prosodic features. In addition to improvement in syllable recognition performance, the system generates the associated sequence of prosody labels as well. This stream of symbolic prosodic information can be useful for many applications - for instance, as additional indexing features for SDR.

One limitation of our approach is that human-annotated prosody labels are required to train the prosody models. While this is not a major constraint for the acoustic-prosodic model (there is enough data to train a 2-way classifier), the prosodic language model is quite parameter-rich and requires much more data to be reliably estimated. Indeed, given a larger amount of annotated data, we would expect higher gains from using the prosody models than we have reported in this paper. More generally, the dependence of such enriched representations as we have discussed in this paper on human-annotated data limits their applicability to domains where such data is available. High-performance unsupervised techniques for automatic annotation of prosodic events are necessary in order to enable these methods to be widely applied.

7. References

[1] A. Sethy, S. Narayanan, and S. Parthasarathy, "A split lexicon approach for improved recognition of spoken names," *Speech Communication*, pp. 1126–1136, September 2006.

[2] M. Larson and S. Eickeler, "Using syllable-based indexing features and language models to improve German spoken document retrieval," in *8th European Conference on Speech Communication and Technology (EUROSPEECH)*, September 2003, pp. 1217–1220.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] S. Ananthakrishnan and S. Narayanan, "Improved speech recognition using acoustic and lexical correlates of pitch accent in a N-best rescoring framework," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

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

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

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

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

[13] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1996.

[14] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1989, pp. 532–535.