A NOVEL ALGORITHM FOR UNSUPERVISED PROSODIC LANGUAGE MODEL ADAPTATION

Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory
Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089
ananthak@usc.edu, shri@sipi.usc.edu

ABSTRACT

Symbolic representations of prosodic events have been shown to be useful for spoken language applications such as speech recognition. However, a major drawback with categorical prosody models is their lack of scalability due to the difficulty in annotating large corpora with prosodic tags for training. In this paper, we present a novel, unsupervised adaptation technique for bootstrapping categorical prosodic language models (PLMs) from a small, annotated training set. Our experiments indicate that the adaptation algorithm significantly improves the quality and coverage of the PLM. On a test set derived from the Boston University Radio News corpus, the adapted PLM gave a relative improvement of 13.8% over the seed PLM on the binary pitch accent detection task, while reducing the OOV rate by 16.5% absolute.

Index Terms— prosody, pitch accent, prosodic language model, unsupervised adaptation, lattice posterior

1. INTRODUCTION

Prosody refers to rhythm, intonation, and lexical stress, and is expressed in speech via modulation of fundamental frequency (F0), prominence patterns, and durational cues such as syllable lengthening and speech rate. These cues occur at the syllable, word, utterance and discourse level, and supplement segment-level information provided by traditional acoustic-phonetic features (e.g. MFCC). However, the high speaker- and context-specific variability exhibited by these acoustic correlates of prosody in combination with their tenuous relationship with the underlying linguistic structure of speech has made it difficult to integrate them with spoken language systems in a systematic fashion.

Categorical representations of prosody offer a solution to this problem by encoding prosodic events using a symbolic alphabet. Tones and Break Indices (ToBI) [1] is one such annotation standard well-known in the community. Symbolic transcription of prosodic events greatly reduces the variability associated with acoustic-prosodic features and makes it easier for automatic learning algorithms to derive relationships between said events and other linguistic entities (e.g. words, syntactic boundaries). Another advantage of some of these schemes is that they are linguistically motivated, and the prosodic transcription is correlated with linguistic entities (e.g. ToBI pitch accent is closely related to syllable stress and word prominence). Hasegawa-Johnson et al. [2] integrated binary pitch accent labels from ToBI-style annotations within an ASR framework to reduce the word error rate (WER) of the system. Their system used joint models of prosodic and spectral features in combination with a prosody-enriched language model. More recently, we have used decoupled prosody models based on categorical representations to re-rank ASR N-best lists [3] and also to enrich ASR lattices [4] for improved recognition performance.

The major disadvantage of symbolic prosody is the cost and effort involved in producing speech corpora with the requisite annotations for learning the linguistic-prosodic models. As a result, few speech corpora have been annotated in this fashion, and those that exist are quite small in comparison to generic speech databases for, say, ASR training. The problem of sparsity, in particular, severely affects the lexical-prosodic model due to its parameter-rich nature, causing a high out-of-vocabulary (OOV) rate on test sets. This limits its usefulness in applications such as prosodic event detection and speech recognition. On the other hand, previous work [5] suggests that the PLM is extremely important for applications that work with symbolic prosody, since it is the glue that binds lexical items to prosodic symbols.

In this paper, we focus on alleviating the sparsity problem suffered by PLMs by proposing a novel algorithm for unsupervised adaptation with a large, unlabeled corpus using a technique reminiscent of confidence-based adaptation in ASR. Different segments of the adaptation set are weighted according to the confidence level assigned by the seed models during automatic prosody labeling. This weighted data is used to adapt the seed PLM. Our experiments indicate that this algorithm results in an improved PLM with significantly

reduced OOV rate. The remainder of this paper is organized as follows: Section 2 contains a summary of the data corpus used in our experiments. Section 3 describes the baseline prosody labeling system, including the prosodic acoustic and language models. In Section 4, we describe the adaptation algorithm in detail. Section 5 summarizes the results of our experiments. Section 6 concludes the paper with a brief discussion of our work and outlines future research possibilities.

2. DATA CORPUS

The Boston University Radio News Corpus (BU-RNC) [6] consists of about 3 hours of read news broadcast speech from 6 speakers (3 male, 3 female) with ToBI-style pitch accent and boundary tone annotations. The entire corpus consisted of 29,573 words, which we split into a training set (14,719 words) and an evaluation set (14,854 words). After eliminating story repetitions from the evaluation set, its useful size was reduced to 10,273 words, which we split into a held-out development set (2,900 words) and a test set (7,373 words). We chose a much smaller training set than usual to simulate real-world situations where very little prosodically annotated data is available, and also to test the efficacy of our algorithm in a data-starved scenario. As before, various types of pitch accents annotated in the BU-RNC were collapsed to binary labels that indicated presence or absence of pitch accents. A total of 7,002 words (47.5%) in the training set carried any type of pitch accent; similarly, 3,471 words (47.0%) in the test set carried a pitch accent.

The adaptation dataset was culled from the WSJ1 (CSR-II) [7] broadcast news speech recognition corpus and consisted of approximately 22,400 utterances (52 hours, 407,000 words). This corpus consists of just the speech data and associated transcriptions, and does not provide symbolic transcription of pitch accents or other prosodic events. The unsupervised algorithms described in the following sections used this corpus to adapt the seed model.

3. BASELINE SYSTEM

The prosodic event detector used in our experiments follows our work in [5], where we proposed a maximum *a-posteriori* (MAP) structure for the prosody recognizer. Thus, our system chooses the sequence of binary pitch accent labels $\mathbf{P}$ that maximizes their posterior probability given the acoustic-prosodic features $\mathbf{A_p}$ and the word sequence $\mathbf{W}$, according to Eq. 1 below.

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

We simplify the above expression by first applying Bayes' rule and then by invoking the assumption that the acoustic-prosodic features are conditionally independent of the lexical evidence, given the sequence of pitch accent labels. Eq. 1 can then be rewritten as follows.

$$\begin{aligned} \mathbf{P}^* &= \arg\max_{\mathbf{P}} p(\mathbf{A_p}, \mathbf{W}|\mathbf{P})p(\mathbf{P}) \\ &\approx \arg\max_{\mathbf{P}} p(\mathbf{A_p}|\mathbf{P})p(\mathbf{W}, \mathbf{P}) \end{aligned} \quad (2)$$

In Eq. 2, the RHS involves two factors - a) the prosodic acoustic model $p(\mathbf{A_p}|\mathbf{P})$, which provides the likelihood of the acoustic-prosodic features given the pitch accent label and b) the PLM $p(\mathbf{W}, \mathbf{P})$, which relates the word sequence to the pitch accent label sequence.

Table 1. Acoustic-prosodic features

Feature	Description
VOWEL_DUR	$\max_{v \in w_i} norm\_dur(v)$
F0AVG_UTT	$\lvert avgF0(w_i) - avgF0(utt)\rvert$
F0RANGE	$maxF0(w_i) - minF0(w_i)$
F0AVG_PAVG	$\lvert avgF0(w_i) - avgF0(w_{i-1})\rvert$
F0AVG_NAVG	$\lvert avgF0(w_i) - avgF0(w_{i+1})\rvert$
F0MAX_PMAX	$\lvert maxF0(w_i) - maxF0(w_{i-1})\rvert$
F0MAX_NMAX	$\lvert maxF0(w_i) - maxF0(w_{i+1})\rvert$
ERMS_AVG	$rmse(w_i)/rmse(utt)$
ERMS_PRMS	$rmse(w_i)/rmse(w_{i-1})$
ERMS_NRMS	$rmse(w_i)/rmse(w_{i+1})$

3.1. Prosodic acoustic model

The acoustic model is implemented as a 25-mixture Gaussian Mixture Model (GMM) with diagonal covariance structure. Since the pitch accent labels are binary (accent vs. no accent), we trained two GMMs, one for each class, using the EM algorithm. Word-level acoustic-prosodic features for training these GMMs are obtained from ASR forced alignment at the word- and phone-level, and are based on previous work on prosody labeling. Table 1 lists a total of 10 features extracted from the F0 track, energy, and vowel duration cues.

3.2. Prosodic language model

The PLM is a joint probability distribution over the word sequence $\mathbf{W}$ and binary pitch accent tags $\mathbf{P}$. We implemented it by creating compound tokens $\mathbf{W}' = (\mathbf{W}, \mathbf{P})$ and training a standard back-off trigram LM with these tokens. This model is trained only on the annotated data from the BU-RNC and will henceforth be referred to as the *seed model* $p_{seed}(\mathbf{W}')$.

3.3. Labeling algorithm

Our word-level pitch accent labeling implementation begins with the construction of a word graph ("sausage") for each test utterance, as shown in Fig. 1. Accented and non-accented variants of a word form the arcs between successive nodes in

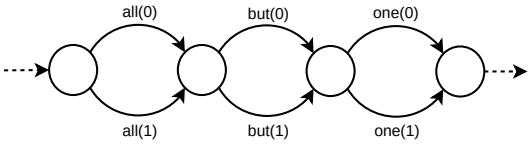

Fig. 1. Word graph with prosodic variants

the graph. Next, we evaluate likelihood scores for the two prosodic variants using the acoustic model and embed these within the corresponding arcs. The graph is then rescored with the seed PLM. Finally, Eq. 2 is implemented using the Viterbi algorithm to determine the best path through the resulting lattice.

4. UNSUPERVISED ADAPTATION

We first describe two naïve techniques for alleviating the high OOV rate of the seed model and compare their performance to that of the seed model as well as that of the proposed scheme.

- *Majority Prediction (MajPred)*: This is a trivially straightforward method in which we associate each word of the adaptation set with the majority prosodic category. In our case, each word is labeled as "not accented", and this labeled corpus is then used to train a PLM that is merged with the seed model to create an adapted model $p_{all0}(\mathbf{W}')$.
- *LM Prediction (LMPred)*: In this method, we train a factored back-off model $p_{seed}(\mathbf{P}|\mathbf{W})$ from the annotated data. The factored structure provides better smoothing for prosody label prediction as compared to the joint model and is described in greater detail in [5]. We use this model to predict prosody labels for the adaptation set. This labeled corpus is then used to train a PLM that is merged with the seed model to generate the adapted model $p_{lmp}(\mathbf{W}')$.

Neither of these simplistic techniques utilizes the discriminatory power provided by the acoustic evidence. Our proposed adaptation algorithm makes use of both the acoustic and lexical models of prosody to train a PLM that is superior to these naïve techniques.

4.1. Proposed algorithm

We begin by setting up the pitch accent detection framework for the unlabeled adaptation data using the acoustic model and the seed PLM as described in Section 3.3. Due to the back-off structure of the PLM, the lattices generated by rescoring the word graph with the seed models no longer retain the original sausage structure.

In the next step, we generate posterior probabilities for each compound token $W' = (W, P)$. This is accomplished

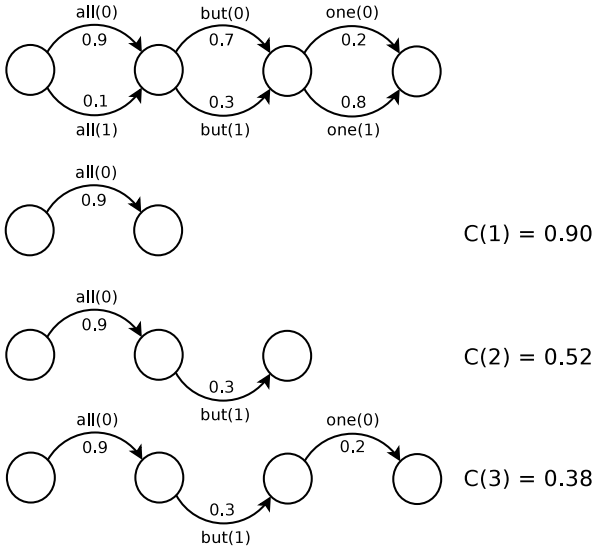

Fig. 2. Obtaining n-gram counts from lattice chunks

by a two-step process: 1) link posteriors $p(l|\mathbf{A_p})$ are computed for each link l in the rescored lattice using a variant of the forward-backward algorithm and 2) links corresponding to the same compound token are collapsed to generate a confusion network identical to the one that was originally created for labeling, except that the arcs in the network now contain compound token posterior probabilities computed from the prosodic acoustic and language models. This technique for generating posteriors and confusion networks is borrowed from minimum word error rate decoding for ASR [8, 9].

The confusion networks with token posteriors are then used to generate fractional (soft) counts for PLM n-gram estimation. Figure 2 illustrates this procedure for a sample fragment of the network. The unigram count of a compound token is set to its posterior probability. For higher order terms (bigrams and trigrams), the count of the term is set to the geometric mean of its constituent posteriors. Thus, terms with high posterior probabilities, which correspond to regions in which the seed models exhibit high confidence, are assigned larger fractional counts and vice-versa. In this way, fractional counts for all possible unigrams, bigrams and trigrams are extracted from the adaptation networks and are used to train a PLM $p_{frac}(\mathbf{W}')$ from the adaptation data. This is merged with the seed PLM to create the adapted model $p_{adapt}(\mathbf{W}')$.

5. EXPERIMENTS AND RESULTS

We split the BU-RNC data into training, development, and testing partitions as described in Section 2. We extracted acoustic-prosodic features using ASR forced alignment information and trained the acoustic model as described in Section 3.1. This model by itself performed with an error rate of 26.6% on the test set. We then trained the seed PLM from the

Table 2. Pitch accent detection results (error rate)

Method	Dev	Test	OOV
Chance	50.0%	47.0%	-
Acoustic	26.4%	26.6%	-
Seed PLM	31.0%	32.0%	21.3%
MajPred PLM	34.0%	34.5%	14.8%
LMPred PLM	29.7%	31.6%	12.8%
Adapted PLM	**24.1%**	**27.6%**	**4.8%**
Combined	18.9%	22.5%	-

annotated data; this model gave an error rate of 32.0% on the test data, significantly worse than the acoustic model due to the high compound token OOV rate of 21.3%.

We experimented with the two naïve methods discussed in Section 4 to generate adapted PLMs and evaluated their performance. The majority prediction (MajPred) scheme performed worse than the seed PLM with an error rate of 34.5%. The token OOV rate was significantly reduced to 14.8% due to inclusion of the adaptation data, but since the two prosodic categories were more or less balanced, majority prediction resulted in a high error rate when used to label the adaptation set, causing poor estimates of the resulting PLM. The LM prediction scheme (LMPred), with an error rate of 31.6% and a token OOV rate of 12.8%, performed slightly better than the seed model. The incremental gain provided by this scheme is due to improved smoothing provided by the factored PLM.

We then implemented the baseline system of Section 3.3 in order to generate lattices for PLM adaptation. After converting the rescored lattices to confusion networks, we extracted fractional unigram, bigram and trigram counts as detailed in Section 4.1. We used the SRILM toolkit [10] to train a PLM using these counts, which we merged with the seed model to generate the adapted model $p_{adapt}(\mathbf{W}')$. The weight of the acoustic model in the initial rescoring step and the PLM merge weight in the final step were jointly optimized on the development set. The error rate for pitch accent labeling with this model was 27.6%, which represented a 4.4% absolute (13.8% relative) reduction in error rate over the seed model. The token OOV rate for this model was only 4.8%. Combining the acoustic model with the adapted PLM resulted in an error rate of 22.5% on the test set. Table 2 summarizes the results of various schemes for the pitch accent detection task.

6. DISCUSSION AND FUTURE WORK

We presented a novel unsupervised adaptation algorithm for improving the quality of PLMs and evaluated its usefulness on the pitch accent detection task. Our proposed scheme results in a 13.8% relative reduction in binary pitch accent labeling error rate and a 16.5% absolute reduction in token OOV rate over the seed model.

One of the major stumbling blocks to using PLMs for speech recognition as presented in the lattice-enrichment framework [4] is the sparsity issue and the concomitant high OOV rate. Our proposed adaptation algorithm significantly reduces the compound token OOV rate and improves the quality of the PLM for prosodic event detection. In the future, we would like to apply the same techniques to improve speech recognition performance.

In this paper, we essentially used the discriminatory power of acoustic evidence to adapt the seed lexical models. Another interesting experiment would be to determine whether token posteriors derived from the confusion networks can be used to adapt the acoustic model for improved performance. This would complete the circle by using knowledge from the seed PLM to adapt the acoustic model.

7. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] S. Ananthakrishnan and S. Narayanan, "Prosody-enriched lattices for improved syllable recognition," in *Proceedings of the International Conference on Spoken Language Processing*, Antwerp, September 2007.

[5] S. Ananthakrishnan and S. Narayanan, "Automatic prosody labeling using acoustic, lexical and syntactic evidence," *to appear in the IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, 2007.

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

[7] *CSR-II (WSJ1) Complete*, Linguistic Data Consortium, Philadelphia, 1994.

[8] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer, Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[9] G. Evermann, "Minimum word error rate decoding," Master's thesis, Cambridge University, 1999.

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