Data-Driven Unsupervised Adaptation of Acoustic-Prosodic Models

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@siпи.usc.edu

Abstract

Training categorical prosody models for spoken language systems requires a significant amount of speech data annotated with the discrete labels of interest. In practice, the difficulty and expense incurred in producing corpora with rich prosodic transcriptions severely limits their integration within applications. In this paper, we explore the possibility of using a large, unlabeled corpus to adapt, in an unsupervised fashion, acoustic-prosodic models trained from a small, human-annotated seed dataset. Our experiments show that the proposed adaptation scheme improves the ability of the acoustic-prosodic model to distinguish between prosodic categories. On a test set derived from the Boston University Radio News Corpus, the adapted models reduced pitch accent detection error rate by 4.4% relative to the seed acoustic-prosodic models trained from the annotated data.

1. Introduction

Symbolic representations of prosody such as those proposed by the ToBI [1] annotation standard have been shown to be useful for spoken language systems, including speech recognition. Models linking discrete prosody labels to linguistic elements (words and syllables) allow us to tap into the parallel stream of information (supplementary to traditional segment-level acoustic-phonetic features, such as MFCCs) contained in speech prosody, in a robust and principled fashion. Hasegawa-Johnson et al. [2] used joint prosodic-phonetic acoustic models and a prosody-enriched language model to improve speech recognition performance. Subsequently, we presented a scheme for using categorical prosody models decoupled from the ASR to rescore N-best lists [3] and to directly enrich ASR lattices with symbolic prosody [4] for improved speech recognition performance.

The key to developing categorical prosody models for spoken language applications is the availability of prosody-annotated speech corpora. Producing such corpora, however, is a laborious and expensive exercise - with the result that such corpora are usually small in size and are available only for select domains. This presents a sparsity issue for training the prosody models. For instance, a Gaussian mixture acoustic-prosodic model (GMM) may require several hundred free parameters to be trained with just a few thousand training samples, causing the model to overfit the training set and preventing generalization to unseen data.

In this paper, we present a technique for unsupervised adaptation of GMM-based acoustic-prosodic models using a much larger, unannotated dataset. Our scheme involves weighting the adaptation data using the seed models (trained from a small, human-annotated corpus), followed by maximum a-posteriori (MAP) adaptation of the seed models using a weighted variant of the expectation-maximization (EM) algorithm. We show that the adapted models outperform the seed models on the binary pitch accent (presence vs. absence) detection task. The remainder of this paper is organized as follows: Section 2 describes the data corpus used in our experiments. Section 3 describes the acoustic and linguistic components of the baseline pitch accent detection system. Section 4 provides an in-depth description of our adaptation scheme. Section 5 summarizes our experimental results and finally, Section 6 concludes the paper with a brief discussion of our findings and outlines future directions for research.

2. Data Corpus

The Boston University Radio News Corpus (BU-RNC) [5] 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. We then performed a 5-fold split of the evaluation set, with 80% (8,218 words) of the data in held-out development sets and 20% (2,055 words) in test sets. These partitions were carried out in such a way that all 5 cross-validation test sets were independent of each other. We chose a much smaller training set than usual to simulate real-world situations where very little prosodically annotated data is available, and 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.

The adaptation dataset was culled from the WSJ1 (CSR-II) [6] 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 [7], where we proposed a maximum a-posteriori
We implemented it by creating compound tokens $W' = \langle W, P \rangle$ and training a standard back-off trigram LM with these tokens. This model is trained only on the annotated data from the BU-RNC. We used the SRILM toolkit [8] to train the prosodic language model.

### 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 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.

![Word confusion network with prosodic variants](image)

### 4. Acoustic-prosodic model adaptation

One straightforward approach to adaptation involves using the seed prosody models to obtain binary pitch accent labels for the adaptation data, and using this automatically annotated data to adapt the seed acoustic models using standard EM-based MAP [9]. However, the seed models are likely to exhibit a higher-than-desirable error rate for pitch accent detection on the unlabeled data, thereby reducing the utility of those data for adaptation. Instead, we propose a soft adaptation approach in which the seed models assign posterior probability scores for prosodic variants of each word. These scores are then used to adapt the seed acoustic-prosodic models using weighted EM-MAP.

#### 4.1. Adaptation data weighting

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

Next, we generate posterior probabilities for each compound token $W' = \langle W, P \rangle$ by a two-step process: 1) link posteriors $p(l|A_p)$ 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 [10, 11]. These posterior probabilities are used as adaptation weights in the modified EM-MAP scheme.

#### 4.2. Weighted EM-MAP

We propose a novel, weighted EM-MAP scheme for soft adaptation of the acoustic-prosodic models using posterior probabilities obtained from the prosodic confusion networks. This method differs from conventional EM training for GMM estimation in that each adaptation sample has a weight associated with it. Samples with larger weights (indicative of high confidence) contribute more to the adaptation process.
samples with low confidence do not have a significant influence on the adapted estimates. A distinctive feature of this approach is that we do not divide the unlabeled data into classes based on confidence scores; rather, all adaptation samples affect both acoustic-prosodic models simultaneously, but to different degrees. The relative influence of each sample on the GMMs is dictated by the external information source, in this case the posterior probability assigned to each sample by the seed models.

We begin by defining a log-likelihood function that incorporates the seed model weights as shown in Eq. 3.

\[
L(\Theta | X, B) = p(X | \Theta, B) = \prod_{i=1}^{N} \sum_{k=1}^{K} \omega_k p_k(x_i | \theta_k, \beta_k) \tag{3}
\]

where \( p_k(x_i | \theta_k, \beta_k) \equiv \mathcal{N}(x_i; \mu, \beta_k^{-1} \Sigma_k) \). This function differs from the traditional log-likelihood function due to integration of the confidence weights \( B \equiv \{\beta_1, \ldots, \beta_N\} \) associated with vector adaptation samples \( X = \{x_1, \ldots, x_N\} \). The rationale behind this modified log-likelihood function is that adaptation samples associated with a large weight “see” a narrow, focused distribution, whereas samples with low confidence weights “see” a diffuse, flat distribution. As we will see, this formulation leads to parameter update equations that emphasize samples with high confidence and vice-versa.

Following the notation of [12], the modified auxiliary function for EM is then given by Eq. 4 (the superscript in \( \Theta^k \) indicates an initial “guess” for the parameter \( \Theta \)).

\[
Q(\Theta, \Theta^k) = \mathcal{E} (\log p(X, Y | \Theta, B) | X, \Theta^k, B) = \sum_{y \in Y} \log p(X, Y | \Theta, B) p(y | X, \Theta^k, B) = \sum_{k=1}^{K} \sum_{i=1}^{N} c_{ik} \log (\omega_k p_k(x_i | \theta_k, \beta_k)) \tag{4}
\]

where \( c_{ik} = p(k | x_i, \Theta^k, \beta_k) = \frac{\omega_k p_k(x_i | \theta_k^k, \beta_k)}{\sum_{j=1}^{K} \omega_j p_j(x_i | \theta_j^k, \beta_k)} \).

Using basic vector and matrix calculus [12], this modified auxiliary function can be maximized w.r.t the unknown parameters to obtain the following maximum-likelihood (ML) update equations for the mixture weights \( \omega_k \), mean vectors \( \mu_k \) and covariance matrices \( \Sigma_k \).

\[
\omega_k' = \frac{1}{N} \sum_{i=1}^{N} c_{ik} \tag{5}
\]

\[
\mu_k' = \frac{\sum_{i=1}^{N} \beta_i c_{ik} x_i}{\sum_{i=1}^{N} \beta_i c_{ik}} \tag{6}
\]

\[
\Sigma_k' = \frac{\sum_{i=1}^{N} \beta_i c_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} c_{ik}} \tag{7}
\]

These modified update equations make intuitive sense: Eq. 6 is the mean of the adaptation samples weighted not only by the mixture occupation likelihoods \( c_{ik} \) as in conventional EM, but also by the confidence weights \( \beta_i \). This suggests that the adaptation samples with large confidence weights influence the estimated GMM mean vectors to a greater extent than samples with low weights. Similarly, Eq. 7 implies that the distributions are focused around the samples with large confidence weights.

While the ML update equations provide intuition on how the confidence weights impact parameter estimation, our task in this paper is to adapt existing seed acoustic-prosodic models using unlabeled data. Maximum a-posteriori (MAP) adaptation is the traditional approach to this problem. Following the approach of [9], we construct a prior distribution for the GMM parameters by assuming the form of a Dirichlet distribution for the mixture weights \( \omega_k \) and a normal-Wishart distribution for the mean vectors \( \mu_k \) and covariance matrices \( \Sigma_k \) as illustrated in Eq. 8.

\[
P(\Theta) \propto \prod_{k=1}^{K} \omega_k^{\lambda_k} |\Sigma_k|^{-1} \omega_k^{-d/2} \exp \left( -\frac{1}{2} (\mu_k - m_k)^T \Sigma_k^{-1} (\mu_k - m_k) \right) \exp \left( -\text{tr}(U_k \Sigma_k^{-1}) \right) \tag{8}
\]

These prior “hyperparameters” are computed using the original (labeled) seed training data in a manner similar to that described in [13]. This leads to the following update equations for weighted EM-MAP.

\[
\omega_k' = \frac{\lambda_k + \sum_{i=1}^{N} c_{ik}}{N + \sum_{k=1}^{K} \lambda_k} \tag{9}
\]

\[
\mu_k' = \frac{\tau_k \mu_k + \sum_{i=1}^{N} \beta_i c_{ik} x_i}{\tau_k + \sum_{i=1}^{N} \beta_i c_{ik}} \tag{10}
\]

\[
\Sigma_k' = \frac{2U_k + S_k + M_k}{2\alpha_k - d + \sum_{i=1}^{N} c_{ik}} \tag{11}
\]

where, for ease of notation, we have defined \( S_k \) and \( M_k \) as follows:

\[
S_k = \sum_{i=1}^{N} \beta_i c_{ik} (x_i - \mu_k)(x_i - \mu_k)^T \tag{12}
\]

\[
M_k = \tau_k (m_k - \mu_k)(m_k - \mu_k)^T \tag{13}
\]

As with standard EM, Eqs. 9, 10, 11 are evaluated iteratively until convergence.

5. Experimental results

The BU-RNC dataset was split into training and evaluation sets as described in Section 2. The evaluation set was further divided into 5 held-out development and cross-validation test sets with 80% of evaluation data (8,218 samples) in the former and 20% (2,055 samples) in the latter. The 5 cross-validation test sets were independent of one another.

We first trained seed acoustic-prosodic GMMs and prosodic language models as described in Section 3. For testing our adaptation scheme in sparse data conditions, we trained more complex seed GMMs with 45 mixtures; we were forced to use diagonal covariance matrices for these models because full-covariance matrices quickly became ill-conditioned as a result of data sparsity. Using our adaptation technique, it is possible to start from these seed diagonal covariance models and train full-covariance models that can better fit the data, possibly leading to improved classification performance as well.

The adaptation data was scored using the seed acoustic-prosodic models and the prosodic language models to generate lattices encoding prosodic variants of each word in the adaptation set. Posterior probabilities (confidence weights) were then
obtained for each adaptation sample by converting the scored lattices to confusion networks as described in Section 4.1. These raw scores were pruned so that only those samples with a large difference between the two class posteriors would be used for adaptation (these samples have a very high likelihood of being labeled correctly by the seed models). The pruned confidence scores were used to adapt the seed models according to the weighted EM-MAP scheme described in Section 4.2. Two parameters were sequentially optimized by evaluating the classification performance of the adapted models on the held-out development data: 1) the weight of the acoustic-prosodic model in Eq. 2 and 2) the pruning threshold for selecting samples from the adaptation set for weighted EM-MAP.

Table 2 summarizes the pitch accent classification error performance of the seed and adapted models averaged across the 5 development and test sets. It is clear from these figures that increasing the number of mixtures from 25 to 45, while maintaining the diagonal covariance structure does not improve classification performance (only 0.5% relative on both held-out and test sets). On the other hand, we note that the adapted models reduce the classification error rate by 4.6% relative to the 45-mixture seed models on the held-out sets and by 4.4% relative on the test sets.

### 6. Discussion

Data sparsity due to lack of large, annotated corpora is a problem encountered by almost all spoken language applications that use categorical representations of prosody. This makes it difficult to learn relationships between prosodic symbols and acoustic-prosodic features or lexical items (syllables and words). In this paper, we presented a novel technique based on a modified EM-MAP algorithm to adapt seed acoustic-prosodic models using a large, unlabeled corpus. This technique permitted us to train full-covariance GMMs, which was not possible with the seed training data due to sparsity. The adapted models provided classification error reduction of 4.4% relative to the seed models on the binary pitch accent classification task.

Our main contribution in this paper was the formulation of a modified EM algorithm for maximum-likelihood or MAP estimation of GMM parameters in the presence of an external knowledge source, which ranks (by assigning numerical weights) the training or adaptation samples in order of their “belongingness” to the model that represents them. This technique is very general and may be applied to arbitrary data.

One limitation of this approach is that the two GMMs are adapted in a generative fashion; while this may likely result in the adapted GMMs better fitting their respective classes, it does not necessarily guarantee better classification performance. Thus, we would like to use the proposed weighted EM approach within a discriminative-based adaptation framework to determine whether it would have a greater impact on classification performance. Our ultimate goal is to use these adapted models within spoken language applications in tandem with our proposed adaptation scheme for prosodic language models [14], beginning with prosody-integrated speech recognition.

### 7. References


