A Mandarin Prosodic Boundary Prediction Model Based on Multi-Source Semi-Supervision

Peiyang Shi\textsuperscript{1,2}, Zengqiang Shang\textsuperscript{1,2}, Pengyuan Zhang\textsuperscript{1,2,*}

\textsuperscript{1}Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, China
\textsuperscript{2}University of Chinese Academy of Sciences, China

\texttt{zhangpengyuan@hccl.ioa.ac.cn}

\section*{Abstract}

High-quality prosodic boundary prediction plays an important role in enhancing speech naturalness and intelligibility in Mandarin text-to-speech tasks. However, traditional methods usually require a large amount of token-level labels, which can hardly be applied in low-resource scenarios. In this paper, to solve this problem, we propose a multi-source semi-supervised model using an HMM to assist BERT-based prosody prediction. Our proposed model implements an alternate training mechanism combining BERT-Prosody and HMM, where BERT takes denoised labels from HMM, providing updated character embedding and weak labels for the latter to form a training cycle. Experimental results show that, compared with baseline methods, the F1 score of our model is raised by $1.01\% / 8.25\%$ respectively at prosodic word/phrase level, approaching the performance of supervised models.

\textbf{Index Terms}: prosodic boundary prediction, Mandarin, semi-supervision, HMM

\section{1. Introduction}

Prosodic boundary prediction, as an essential component in a cascading Mandarin text-to-speech pipeline, influences the naturalness and intelligibility of synthesized speech directly. As shown in Figure 1, a prosodic structure is a three layer tree structure, and Mandarin prosodic boundary labels are composed of prosodic words (PW), prosodic phrases (PPH), and intonation phrases (IPH)\cite{1}.

With the development of speech synthesis technology, large numbers of studies have been carried out on prosodic boundary prediction in recent years. Research in the early stages mainly focused on the mining of language features\cite{2–9} and statistical models such as the maximum entropy model\cite{10} and CRF\cite{11}. Afterwards, methods based on deep learning became gradually popular and achieved good performance. Zheng et al.\cite{12} introduced the BLSTM-CRF model; Lu et al.\cite{13} used self-attention combined with word segmentation auxiliary tasks to predict prosodic boundaries; Pan\cite{14} used multi-task learning, which combines a large number of language features to enhance prediction accuracy. Inspired by the success of pre-trained language models (PLM), some researchers\cite{15–20} have demonstrated that PLM can help the model capture rich semantic and syntactic knowledge, improving the performance of text-to-speech (TTS) front-end related tasks.

However, for sequence labeling tasks such as prosodic boundary prediction, existing methods require a large amount of high-quality labeled data. Because prosody labeling is a challenging task that requires abundant professional knowledge, it is necessary to explore effective methods in low-resource scenarios. Learning prosodic boundary prediction models from small datasets is a challenging problem. So far, limited studies have taken place, in which either unsupervised or semi-supervised methods are employed for prosodic prediction. Ananthakrishnan\cite{21} proposed an unsupervised algorithm for prosodic event detection, whose algorithm was based on clustering techniques to make use of acoustic and syntactic cues. Liu\cite{22} applied a co-training algorithm to automatic prosodic event detection for better sample selection to help prediction. These methods are suitable for English prosodic labeling but do not work well if they are replicated in similar tasks for Mandarin. In addition, Wang\cite{23} proposed a self-training method to solve the problem of data insufficiency, however generating excessive noise, which resulted in model overfitting.

In this paper, we propose a multi-source semi-supervised model based on BERT-Prosody. The training stage of BERT relies on an Augmented HMM that extracts high-confidence labels from weak labels of multi-source. The Augmented HMM is responsible for label denoising, which models true labels as hidden variables and infers them from the observed noisy labels in an unsupervised way\cite{24–26}. In addition, we also propose an alternate training mechanism combining HMM and BERT-Prosody. The training process is a two-stage iterative loop: In the first stage, the high-confidence labels generated by HMM are used to fine-tune BERT-Prosody. Then in the second stage, the fine-tuned BERT-Prosody updates the input embedding of HMM and generates new weak labels to replace the worst ones in those old labels for the next iteration.

Our contributions include:

1. Proposing a semi-supervised training method for Mandarin prosody prediction, which can effectively utilize the information of various weak labels and reduce noise.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{example.png}
\caption{Example of mandarin prosody structure. (Note: PW-Prosodic Word, PPH-Prosodic Phrase, IPH-Intonational Phrase)}
\end{figure}
2. Designing an alternate training mechanism that trains the Augmented HMM and BERT-Prosody sequentially in cycles, utilizing the outputs of each other to optimize model performance.

3. Suppressing over-fitting problems and bringing about 1.01%/8.25% in F1-score on PW/PPH respectively, compared with the baseline method.

2. A semi-supervised prosody prediction model

In this section, we will introduce the BERT-based prosody prediction model, illustrate the overall training process of the model, explain the Augmented HMM.

2.1. Training Procedure

We train the Augmented HMM and the BERT-Prosody model in an alternating method. The training process is an iterative loop consisting of two stages. Firstly, the Augmented HMM extracts high confidence labels from weak labels for BERT-Prosody, and then BERT-Prosody improves the enhancement effect of HMM by updating weak labels and character embedding for HMM.

![Figure 2: The illustration of the training method](image)

Figure 2 depicts the entire training process. At each iterative, the Augmented HMM first infers a set of denoised labels \( y^{(t)} \) from a set of weak labels \( x^{(t)} \), which is generated from \( K \) different weak information sources \( F_{1,K} \). Here, weak information sources are prosody prediction models pre-trained on a small data set. Then, this set of labels \( y^{(t)} \) is used to train the BERT-Prosody model.

After the BERT-Prosody model receives those denoised labels, it replaces the set of labels of the most severe noise\(^1\) in the current weak labels with the output labels \( y^{(t)}_{BERT} \), and updates the BERT Embedding required by the HMM.

2.2. The Augmented HMM

The Augmented HMM is a neutralized HMM that models true entity labels as hidden variables and infers them from the observed noisy labels, thereby removing noise from weak labels. Traditional HMMs use fixed transition and emission matrices for training and inference with constant values. In other words, they do not change along with contexts. Augmented HMM uses the character vector of BERT Embedding as input, treating the transition matrix and the emission matrix as the output of a neural network, and thus the Augmented HMM makes better use of the context information.

\[ A(t) = \text{Softmax}(\text{reshape}(FFL(e(t)))) \]

\[ B(t) = \text{Softmax}(\text{reshape}(FFL(e(t)))) \]

where

\[ \text{Softmax}(n) = \frac{\exp(n)}{\sum_i \exp(n_i)} \]

2.2.1. the Augmented HMM structure

Figure 3 shows the brief structure of the Augmented HMM. We assume that there are \( K \) weak sources \( F_k \in \{1, 2, 3...K\} \) generating weak label set \( x^{(t)} \) at time step \( t \). We note the label set as \( L \), then \( x^{(t)} \in L \). The state \( s^{(t)} \) in the figure is the hidden state of the HMM, i.e. the real label we want to derive. \( A(t) \) is the transition matrix, the size is \( |L| \times |L| \), where \( A_{i,j} = p(s^{(t)} = j \mid s^{(t-1)} = i) \) and \( B(t) \) is the emission matrix, the size is \( |L| \times |L| \), where \( B_{i,k} = p(x_k^{(t)} = j \mid s^{(t)} = i) \). \( A(t) \) and \( B(t) \) are generated by \( e(t) \) (char vector generated by BERT) through simple feedforward linear layers (FFL), respectively. Then we use the Softmax function along the label direction to ensure that the sum of the probability distribution inside the matrix is 1:

\[ A(t) = \text{Softmax}(\text{reshape}(FFL(e(t)))) \]

\[ B(t) = \text{Softmax}(\text{reshape}(FFL(e(t)))) \]

2.2.2. Parameter Estimation

Traditional HMM learns the parameters with expectation maximization (EM, also called as Baum-Welch) algorithm. In the Augmented HMM, the transition and emission matrices are generated by FFL rather than learned. Thus, we train these FFL directly, combining the traditional EM method with gradient descent. In the expectation step (E-step), we calculate the expected log-likelihood of the complete data:

\[ Q(\theta, \theta^{old}) = \mathbb{E}_{l \mid (\theta)}[Q(\theta^{old})] \]

Different from the traditional EM algorithm, in the maximization step (M-step), we calculate its gradient descent, and then combine with the gradient descent algorithm to update iteratively.

\[ \nabla \theta_{HMM} = \frac{\partial Q(\theta_{HMM}, \theta^{old}_{HMM})}{\partial \theta_{HMM}} \]

\(^1\)We use the F1 value of weak source to describe the noise level.
## 2.2.3. Parameter Initialization

Different from traditional HMM initializes the transition matrix $A^*$ and emission matrix $B^*$ with statistical data, the Augmented HMM initializes the matrices $A^{(t)}$ and $B^{(t)}$ by pretraining the FFL before the training of the HMM. The pretraining of FFL is by minimizing the mean squared error (MSE) loss between the target statistics and the output of the FFL:

$$
loss = \frac{1}{T} \sum_i \| A^* - A^{(t)} \|^2_F + \| B^* - B^{(t)} \|^2_F
$$

As for the statistics data $A^*$ and $B^*$ based on the observed data statistics, we initialize each row of the state transition matrix as a Dirichlet distribution based on the count $\delta$ of transitions between the observed classes:

$$
A^*_{i,j} = \text{Dirichlet}(\delta_{i,j})
$$

The emission matrix for each weak source specifies the message start value. Assuming that we can provide a rough estimate of recall $r_{j,k}$ and precision $p_{j,k}$ for weak sources $k$ on label $j$, the initial value for the parameters of the emission matrix is expressed as:

$$
B^*_{i,j,k} = \begin{cases} 
  r_{j,k} & i = j \\
  \frac{(1 - r_{j,k})(1 - p_{j,k})}{n_{class} - 1} & i \neq j
\end{cases}
$$

### 2.2.4. Decoding

Once the parameters of the HMM model are determined, we can infer the estimation of the true label $y^*$ from those noisy weak labels $x^*_1:T$. These results can be calculated through Viterbi decoding algorithm [27].

## 2.3. BERT-Prosody

The raw input text is first encoded with a well-prepared Chinese BERT [28], which is composed of a stack of Transformer blocks and pretrained with an adequate amount of Mandarin text data. Benefiting from this pre-training, the model is assumed to be capable of capturing rich contextual and semantic information of Mandarin language. Figure 4 shows the structure of BERT-Prosody, we use the cased-version of the base BERT model (trained on Wikipedia and news text) and perform task-tuning. That is, for each character in the input text, a feature vector will be generated. The prosody prediction layer is built with MLP and a Softmax layer.

![Figure 4: Architecture of the BERT-Prosody](https://huggingface.co/bert-base-uncased)

We train the model by minimizing the cross-entropy error between the output labels of the Augmented HMM $y$ and the output labels of the BERT-Prosody $y^*$. During training, we choose different learning rates for the BERT layer and the linear layer to achieve a better training result.

$$
\theta = \text{argmin} \sum_{i=1}^n E[\text{loss}(y^*, y)]
$$

## 3. Experiment

To verify the effectiveness of the proposed method, we take Self-training [23] as the baseline model. We also study the alternating training process to evaluate the improvement of the model performance by alternating training.

### 3.1. Datasets

Due to the fact that finding a large-scale public dataset for Mandarin prosodic boundary prediction is difficult, we collect a corpus of size 20,000 for prosodic structure prediction (10,000 of the dataset are from the data-baker public datasets, and the other 10,000 are collected by ourselves). To demonstrate the feasibility of our semi-supervised method, we randomly select 500 sentence from the dataset as the training set and 2000 sentence for testing. The remaining data are treated as unlabeled data in semi-supervised training. Statistical information of the entire dataset is shown in Table 1. For the automatic evaluation metrics, we report the F1 score of the PW, PPH of the prosodic structure.

### 3.2. Data Processing

According to statistical results, an IPH often appears somewhere strongly related to a certain type of punctuation in sentences. Hence, in practical applications, we often regard such type of punctuation as a sign of intonation phrases. Therefore, in this experiment, we focus on the prediction of PW and PPH boundaries.

![Figure 5: Example of prosody tag sequence processing](https://www.data-baker.com/open-source.html)

First, we extract K mixed label sequences from weak sources and degrade IPH into PPH labels due to the reason above. Then our model is built in character level, using #0 to denote a blank label inside the prosodic word. Finally, we can obtain label sequences of PW and PPH. The transcription process is shown in Figure 5.
3.3. Weak Sources

We use the well-known CRF model, the BLSTM-CRF model and the BERT-based prosodic boundary prediction model as weak source:

1. CRF: We implement our CRF model with the open source tool CRF++. The input language features include word segmentation information, POS tag, and word length.

2. BLSTM-CRF: Refer to the paper [12] and input the information of word, POS tag, and word length into the BLSTM-CRF network.

3. BERT-prosody: BERT encoder with only one linear layer and one Softmax layer added.

The above models are trained with the known data set with 500 pieces of labeled data. After convergence, these three models are applied to the remaining unlabeled data to generate three weak labels respectively. The F1 performances on the test set are shown in Table 2.

<table>
<thead>
<tr>
<th>PW</th>
<th>ACC</th>
<th>REC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>0.8799</td>
<td>0.8960</td>
<td>0.8879</td>
</tr>
<tr>
<td>BLSTM-CRF</td>
<td>0.8817</td>
<td>0.9069</td>
<td>0.8890</td>
</tr>
<tr>
<td>BERT-Prosody</td>
<td>0.8404</td>
<td>0.9522</td>
<td>0.8928</td>
</tr>
</tbody>
</table>


3.4. Main Result

First, we introduce two bounds in our experiments: 1) the result of BERT-Prosody trained in a fully supervised manner as the upper bound, 2) the result of BERT-Prosody trained with 500 datasets as the lower bound. At the same time, we use the Self-Training method [23] as the baseline for comparison. The experimental results are as follows.

<table>
<thead>
<tr>
<th>PW</th>
<th>ACC</th>
<th>REC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised BERT</td>
<td>0.9245</td>
<td>0.9627</td>
<td>0.9433</td>
</tr>
<tr>
<td>500 datasets BERT</td>
<td>0.8404</td>
<td>0.9522</td>
<td>0.8928</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.9016</td>
<td>0.9470</td>
<td>0.9237</td>
</tr>
<tr>
<td>Semi-Supervision BERT</td>
<td>0.9139</td>
<td>0.9547</td>
<td>0.9338</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PPH</th>
<th>ACC</th>
<th>REC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised BERT</td>
<td>0.8567</td>
<td>0.9110</td>
<td>0.8830</td>
</tr>
<tr>
<td>500 datasets BERT</td>
<td>0.6949</td>
<td>0.8589</td>
<td>0.7683</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.7156</td>
<td>0.8672</td>
<td>0.7842</td>
</tr>
<tr>
<td>Semi-Supervision BERT</td>
<td>0.8508</td>
<td>0.8832</td>
<td>0.8667</td>
</tr>
</tbody>
</table>

Table 3 shows the model performance at PW and PPH levels respectively. We find that our training method outperforms the baseline model. It is highly possible that on a small dataset like we have prepared, the baseline model Self-Training falls into the situation of overfitting. Moreover, it is noticeable that at PPH level, performance of the baseline method declines even further. Compared to the baseline methods, our semi-supervised model shows satisfying performances, with the F1 scores at the prosodic word and prosodic phrase levels improved by 1.01% and 8.25% respectively. It can also be found that the results of our model trained with only 500 labeled data are very approach to the fully supervised model, which further demonstrates the superiority of this design.

3.5. Analysis of Alternate-Training

The details of the alternate training process are demonstrated in Figure 6. The results show that the F1 score of HMM at either PW/PPH level improves greatly after the first round of training, indicating that HMM benefits from the updated information provided by BERT-Prosody. At the same time, the F1 score of BERT-Prosody at either PW/PPH level also has an improvement after the first round of training. In addition, the F1 score of BERT-Prosody at PPH level continues to improve in several subsequent rounds. This experimental result illustrates that the alternate training of HMM and BERT-Prosody is able to optimize the performance of prosody prediction and make the model especially effective on complex tasks such as prediction at PPH level.

4. Conclusions

This paper proposes a multi-source semi-supervised model based on BERT-Prosody and the Augmented HMM. The Augmented HMM is used to extract high-confidence labels from weak labels for the training of BERT-Prosody. The augmented HMM and BERT-Prosody are trained alternately during the training process: the augmented HMM provides high-confidence labels for BERT-Prosody, and BERT-Prosody provides updated BERT embedding and weak labels for HMM. This model can avoid overfitting problems when training on small datasets and significantly improve the final prosody prediction results. In the future, we will further explore whether syntactic-related knowledge could be used to increase the per-
formance of HMM augmented models and examine the performance of such an approach when applied to larger datasets.

5. References


