Minimum Boundary Error Training for Automatic Phonetic Segmentation

Jen-Wei Kuo and Hsin-Min Wang

Institute of Information Science
Academia Sinica, Taipei, Taiwan, Republic of China
(rogerkuo, whm)@iis.sinica.edu.tw

Abstract

Annotated speech corpora are indispensable to various areas of speech research. In this paper, we present a novel discriminative training approach for HMM-based automatic phonetic segmentation. The objective of the proposed minimum boundary error (MBE) discriminative training approach is to minimize the expected boundary errors over a set of phonetic alignments represented as a phonetic lattice. This approach is inspired by the recently proposed minimum phone error (MPE) training algorithm for automatic speech recognition. To evaluate the MBE training approach, we conducted automatic phonetic segmentation experiments on the TIMIT acoustic-phonetic continuous speech corpus. The MBE-trained HMMs can identify 79.75% of human-labeled phone boundaries within a tolerance of 10 ms, compared to 71.23% identified by the conventional ML-trained HMMs. Moreover, by using the MBE-trained HMMs, only 7.89% of automatically labeled phone boundaries have errors larger than 20 ms.

Index Terms: minimum boundary error, automatic phonetic segmentation, HMM, forced alignment

1. Introduction

The development of speech technology has relied heavily on corpus-based methodologies. One of the most important and useful annotations is transcription and segmentation at the phonetic level. In speech recognition, the use of Hidden Markov Models (HMMs) has made manual phonetic segmentation unnecessary, because the HMM training is an averaging process that tends to smooth segmentation errors. However, some researchers believe that speech recognition would benefit from more precise segmentation in training and recognition. For example, it is essential that model bootstrapping should have better initial estimates of the HMM parameters so that the local maximum is as close as possible to the global maximum of the objective function. On the other hand, in recent years, increased attention has been given to the data-driven, concatenation-based TTS synthesis because of its high degree of naturalness and fluency. Both the development of concatenative acoustic unit inventories and the statistical training of data-driven prosodic models require a speech database that is precisely segmented. In the past, synthesis has relied on manual segmentation; however, this is extremely time consuming and costly. To reduce the human effort and speed up the labeling process, many attempts have been made to utilize automatic phonetic segmentation approaches to provide initial phonetic segmentation for subsequent manual segmentation and verification, e.g., dynamic time warping (DTW) [1], methods that utilize specific features and algorithms [2], HMM-based Viterbi forced alignment [3], and discriminative training schemes [4]. These approaches not only save time and money, but also make possible the rapid adaptation of a TTS synthesis system to new voices and languages.

The most common method of automatic phonetic segmentation is to adapt an HMM-based phonetic recognizer to align a phonetic transcription with a speech utterance. Empirically, phone boundaries obtained in this way should contain few serious errors, since HMMs in general capture acoustic properties of phones; however, small errors are inevitable because HMMs are not sensitive enough to detect changes between adjacent phones [4]. Unfortunately, even a small segmentation error may produce an audible error in synthetic speech. To improve the discriminability of HMMs for automatic phonetic segmentation, we propose a novel discriminative training approach that applies a minimum boundary error criterion, instead of the maximum likelihood criterion used in conventional training approaches.

The remainder of this paper is organized as follows. Section 2 describes the proposed minimum boundary error discriminative training scheme in detail. Section 3 presents the experiment results. Finally, in Section 4, we present our conclusions and indicate the direction of our future work.

2. Minimum boundary error training

Given a training set of observation sequences \( O = \{ O_1, \ldots, O_R \} \), the MBE criterion for acoustic model training tries to minimize the expected boundary errors in the sequences. Therefore, according to the MBE criterion, the objective function can be defined as:

\[
F_{\text{MBE}} = \sum_{r=1}^{R} \sum_{S} \Phi_r \cdot P(S_r | O_r) \cdot ER(S_r),
\]

where \( \Phi_r \) is a set of various possible phonetic alignments for the training observation utterance \( O_r \); \( S_r \) is one of the hypothesized alignments in \( \Phi_r \); \( P(S_r | O_r) \) is the posterior probability of alignment \( S_r \), given the training observation sequence \( O_r \); and \( ER(S_r) \) denotes the “boundary error” of \( S_r \) compared with the manually labeled phonetic alignment in the canonical transcription. For each training observation sequence \( O_r \), \( F_{\text{MBE}} \) gives the weighted average boundary error of all hypothesized alignments. For simplicity, we assume the prior probability of alignment \( S_r \) is uniformly
distributed, and the likelihood \( P(O_r | S_r) \) of alignment \( S_r \) is governed by the acoustic model parameter set \( \lambda \). Therefore, Eq.(1) can be rewritten as:

\[
F_{\text{MPE}} = \sum_{r=1}^{R} \sum_{c \in \Phi} P_c(O_r | S_r) \alpha \frac{1}{\sum_{c \in \Phi} P_c(O_r | S_r)} \text{ER}(S_r),
\]

(2)

where \( \alpha \) is a scaling factor that prevents the denominator \( \sum_{c \in \Phi} P_c(O_r | S_r) \) being dominated by only a few alignments. If \( \alpha \) is set to zero, all the hypotheses are equally weighted. Accordingly, the optimal parameter set \( \lambda^* \) can be estimated by minimizing the objective function defined in Eq.(2), i.e.,

\[
\lambda^* = \arg \min_{\lambda} \sum_{r=1}^{R} \sum_{c \in \Phi} P_c(O_r | S_r) \alpha \frac{1}{\sum_{c \in \Phi} P_c(O_r | S_r)} \text{ER}(S_r).
\]

(3)

The boundary error \( \text{ER}(S_r) \) of the hypothesized alignment \( S_r \) can be calculated as the sum of the boundary errors of the individual phones in \( S_r \), i.e.,

\[
\text{ER}(S_r) = \sum_{q \in S_r} \sigma(r),
\]

(4)

where \( q \) is a phone involved in \( S_r \); \( \sigma(r) \) is a phone boundary error function defined as,

\[
\sigma(r) = 0.5 \times |s_q - s_q'| + |e_q - e_q'|,
\]

(5)

where \( s_q \) and \( e_q \) are, respectively, the hypothesized start time and end time of phone \( q \); \( s_q' \) and \( e_q' \) correspond to the manually labeled start time and end time, respectively. Since \( \Phi \) contains a huge number of hypothesized phonetic alignments, it is impractical to sum the boundary errors directly without first pruning some of the alignments. For efficiency, it is suggested that a reduced hypothesis space, e.g., an \( N \)-best list [5] or a lattice (or graph) [6], should be used. However, an \( N \)-best list often contains too much redundant information, e.g., two hypothesized alignments can be very similar. In contrast, as illustrated in Figure 1, a phonetic lattice is more effective because it only stores the lattice; and is the weighted average boundary error of a given Gaussian mixture \( m \) thus derived can be expressed, respectively, as:

\[
\mu_m = \frac{\theta_m^{\text{MPE}}(O) + D_m \overline{\mu}_m}{\gamma_m^{\text{MPE}} + D_m},
\]

(7)

and

\[
\Sigma_m = \frac{\theta_m^{\text{MPE}}(O^2) + D_m \overline{\Sigma}_m + \overline{\mu}_m \overline{\mu}_m^T}{\gamma_m^{\text{MPE}} + D_m}.
\]

(8)

In Eqs. (7) and (8), \( D_m \) is a per-mixture level control constant that ensures all the variance updates are positive; \( \overline{\mu}_m \) and \( \overline{\Sigma}_m \) are the current mean vector and covariance matrix, respectively; and \( \theta_m^{\text{MPE}}(O) \), \( \theta_m^{\text{MPE}}(O^2) \), and \( \gamma_m^{\text{MPE}} \) are, respectively, statistics defined as:

\[
\theta_m^{\text{MPE}}(O) = \sum_t \sum_{q \in \Phi} \sum_{i \in \text{tr}} \gamma_q^{\text{MPE}} \gamma_{q_m}(t) \gamma_{o_i}(t),
\]

(9)

\[
\theta_m^{\text{MPE}}(O^2) = \sum_t \sum_{q \in \Phi} \sum_{i \in \text{tr}} \gamma_q^{\text{MPE}} \gamma_{q_m}(t) \gamma_{o_i}(t) \gamma_{o_i}(t)^T,
\]

(10)

and

\[
\gamma_m^{\text{MPE}} = \sum_t \sum_{q \in \Phi} \sum_{i \in \text{tr}} \gamma_q^{\text{MPE}} \gamma_{q_m}(t).
\]

(11)

In Eqs. (9), (10), and (11), \( \gamma_{q_m}(t) \) is the occupation probability for mixture \( m \) on \( q \), \( o_i(t) \) is the observation vector at time \( t \), and \( \Phi^\text{lat} \) represents the lattice for sentence \( O_r \). \( \gamma_q^{\text{MPE}} \) is computed by

\[
\gamma_q^{\text{MPE}} = \gamma_q^{\text{MPE}} \overline{\eta}_m - \overline{\eta}_q,
\]

(12)

where \( \overline{\eta}_q \) is the occupation probability of phone arc \( q \), also referred to as its posterior probability; \( \overline{\eta}_m \) is the weighted average boundary error of all the hypothesized alignments in the lattice; and \( \overline{\eta}_q \) is the weighted average boundary error of auxiliary function [8] and has been applied in the minimum phone error (MPE) discriminative training approach [9] for ASR, can be adapted to solve Eq.(6). The re-estimation formulae for the mean vector \( \mu_m \) and the diagonal covariance matrix \( \Sigma_m \) of a given Gaussian mixture \( m \) thus derived can be expressed, respectively, as:

\[
\mu_m = \frac{\theta_m^{\text{MPE}}(O) + D_m \overline{\mu}_m}{\gamma_m^{\text{MPE}} + D_m},
\]

(7)

and

\[
\Sigma_m = \frac{\theta_m^{\text{MPE}}(O^2) + D_m \overline{\Sigma}_m + \overline{\mu}_m \overline{\mu}_m^T}{\gamma_m^{\text{MPE}} + D_m}.
\]

(8)

In Eqs. (7) and (8), \( D_m \) is a per-mixture level control constant that ensures all the variance updates are positive; \( \overline{\mu}_m \) and \( \overline{\Sigma}_m \) are the current mean vector and covariance matrix, respectively; and \( \theta_m^{\text{MPE}}(O) \), \( \theta_m^{\text{MPE}}(O^2) \), and \( \gamma_m^{\text{MPE}} \) are, respectively, statistics defined as:

\[
\theta_m^{\text{MPE}}(O) = \sum_t \sum_{q \in \Phi} \sum_{i \in \text{tr}} \gamma_q^{\text{MPE}} \gamma_{q_m}(t) \gamma_{o_i}(t),
\]

(9)

\[
\theta_m^{\text{MPE}}(O^2) = \sum_t \sum_{q \in \Phi} \sum_{i \in \text{tr}} \gamma_q^{\text{MPE}} \gamma_{q_m}(t) \gamma_{o_i}(t) \gamma_{o_i}(t)^T,
\]

(10)

and

\[
\gamma_m^{\text{MPE}} = \sum_t \sum_{q \in \Phi} \sum_{i \in \text{tr}} \gamma_q^{\text{MPE}} \gamma_{q_m}(t).
\]

(11)

In Eqs. (9), (10), and (11), \( \gamma_{q_m}(t) \) is the occupation probability for mixture \( m \) on \( q \), \( o_i(t) \) is the observation vector at time \( t \), and \( \Phi^\text{lat} \) represents the lattice for sentence \( O_r \). \( \gamma_q^{\text{MPE}} \) is computed by

\[
\gamma_q^{\text{MPE}} = \gamma_q^{\text{MPE}} \overline{\eta}_m - \overline{\eta}_q,
\]

(12)

where \( \overline{\eta}_q \) is the occupation probability of phone arc \( q \), also referred to as its posterior probability; \( \overline{\eta}_m \) is the weighted average boundary error of all the hypothesized alignments in the lattice; and \( \overline{\eta}_q \) is the weighted average boundary error of

\[
\text{Figure 1: An illustration of the phonetic lattice for the speech utterance “where were they?”}. The lattice can be generated by performing a beam search using some pruning techniques.
\]
the hypothesized alignments in the lattice that contain arc \( q \).
Note that the term \( \eta_{avg}^r - \eta_q^r \) reflects the difference between
the weighted average boundary error of all the alignments in
the lattice and that of the alignments containing arc \( q \). When
\( \eta_{avg}^r \) equals \( \eta_q^r \), phone arc \( q \) makes no contribution to MBE
training. However, when \( \eta_{avg}^r \) is larger than \( \eta_q^r \), i.e., phone
arc \( q \) generates fewer errors than the average, then \( q \) makes a
positive contribution. Conversely, if \( \eta_{avg}^r \) is smaller than \( \eta_q^r \),
\( q \) makes a negative contribution. The discriminative ability of
the MBE training approach is thus shown. \( \gamma_q^r \), \( \eta_{avg}^r \), and \( \eta_q^r \)
are computed by

\[
\gamma_q^r = \frac{\sum_{S_i} \phi_q^{wq} \eta_q^{wq} S_i P(O_r | S_i)^\alpha}{\sum_{S_i} \phi_q^{wq} P(O_r | S_i)^\alpha},
\]

\[
\eta_{avg}^r = \frac{\sum_{S_i} \phi_q^{wq} \eta_q^{wq} S_i P(O_r | S_i)^\alpha}{\sum_{S_i} \phi_q^{wq} P(O_r | S_i)^\alpha},
\]

and

\[
\eta_q^r = \frac{\sum_{S_i} \phi_q^{wq} \eta_q^{wq} S_i P(O_r | S_i)^\alpha}{\sum_{S_i} \phi_q^{wq} P(O_r | S_i)^\alpha},
\]

respectively, where \( \lambda \) is the current set of parameters. The
above three quantities can be calculated efficiently by applying
dynamic programming to the lattice.

### 2.2 I-smoothing update

To improve the generality of MBE training, the I-smoothing
technique [9] is employed to provide better parameter
estimates. This technique can be regarded as interpolating the
MBE and ML auxiliary functions according to the amount of
data available for each Gaussian mixture. The updates for the
mean vector \( \mu_m \) and the diagonal covariance matrix \( \Sigma_m \) thus become:

\[
\mu_m = \frac{\theta_m^{MBE}(O) + D_m \mu_m + \tau_m \theta_m^{ML}(O)}{\gamma_m^{MBE} + D_m + \tau_m},
\]

and

\[
\Sigma_m = \frac{\theta_m^{MBE}(O) + D_m \Sigma_m + \tau_m \theta_m^{ML}(O)}{\gamma_m^{MBE} + D_m + \tau_m} - \mu_m \mu_m^T,
\]

respectively, where \( \tau_m \) is also a per-mixture level control
constant, and

\[
\gamma_m^{ML} = \sum_{r=1}^{R} \gamma_m^{ML}(t),
\]

\[
\theta_m^{ML}(O) = \sum_{r=1}^{R} \gamma_m^{ML}(t) \alpha_r(t),
\]

\[
\theta_m^{ML}(O^2) = \sum_{r=1}^{R} \sum_{i=1}^{T} \gamma_m^{ML}(t) \alpha_r(t) \alpha_r(t)^\top.
\]

### 3. Experiments

#### 3.1 Experiment setup

TIMIT (The DARPA TIMIT Acoustic-Phonetic Continuous
Speech Corpus) [10], a well-known read speech corpus with
manual acoustic phonetic labeling, has been widely used for
the evaluation of automatic speech recognition and phonetic
segmentation. TIMIT contains a total of 6,300 sentences,
comprised of 10 sentences spoken by each of 630 speakers
from 8 major dialect regions in the United States. The TIMIT
suggested training and testing sets contain 462 and 168
speakers, respectively. We discard utterances with phones
shorter than 10 ms. The resulting training set contains 4,546
sentences, with a total length of 3.87 hours, while the testing
set contains 1,646 sentences, with a total length of 1.41 hours.

The acoustic models consist of 50 context-independent
phone models, each represented by a 3-state continuous
density HMM (CDHMM) with a left-to-right topology.
Each frame of the speech data is represented by a 39-
dimensional feature vector comprised of 12 MFCCs and log energy,
and their first and second differences. The frame width
is 20 ms and the frame shift is 5 ms. Utterance-based cepstral
variance normalization (CVN) is applied to all the training and
testing speech.

#### 3.2 Experiment results

The acoustic models were first trained on the training speech
according to the human-labeled phonetic transcriptions and
boundaries by the Baum-Welch algorithm using the ML
criterion with 10 iterations. Then, the MBE discriminative
training approach was applied further to manipulate the models.
The scaling factor \( \alpha \) in Eq.(2) was set to 0.1 and the I-
smoothing control constant \( \tau_m \) in Eqs.(16) and (17) was set to
20 for all mixtures. The results are shown in Figure 2. The line
with triangles in the figure indicates the expected FER (frame
error rate) calculated at each iteration of the training process.
Clearly, the descending trend satisfies the training criterion.
We have explored the use of the minimum boundary error (MBE) criterion in the discriminative training of acoustic models for automatic phonetic segmentation. The underlying characteristics of MBE training have been investigated, and its superiority over conventional ML training has been verified by experiments. Naturally, the more accurate phonetic segmentation obtained by the MBE-trained models is very useful for subsequent manual verification or further boundary refinement using other techniques. The MBE training method is not difficult to implement, in particular some discriminative training tools, such as MPE, have been included in HTK.

In addition to applying the MBE criterion to the training of acoustic models, we have applied it to discriminative feature training. The preliminary experiment results indicate that feature-based MBE training is more effective than model-based MBE training. The segmentation accuracy could be improved by integrating the feature-based and model-based MBE training procedures. Moreover, a new decoding algorithm based on the minimum boundary error criterion is also under development. It is hoped that phone boundaries can be located more accurately by running a second pass search using the minimum boundary error criterion on the lattice generated by a first pass conventional search. In our current implementation, the phone boundary error function, defined in Eq.(5), is calculated in the time frame unit for efficiency. However, more accurate segmentation may be achieved by calculating boundary errors in actual time sample marks.

5. Acknowledgements
This work was funded by the National Science Council of the Republic of China under Grant: NSC94-2213-E-001-021.

6. References

Table 1: The percentage of phone boundaries correctly placed within different tolerances with respect to their associated manually labeled phone boundaries.

<table>
<thead>
<tr>
<th>Boundary Distance</th>
<th>ML10</th>
<th>ML20</th>
<th>ML10 + MBE10</th>
<th>ML10 + MBE10 + I-smoothing</th>
<th>Absolute improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5ms</td>
<td>46.69</td>
<td>46.95</td>
<td>58.35</td>
<td>58.48</td>
<td>11.53</td>
</tr>
<tr>
<td>&lt;10ms</td>
<td>71.10</td>
<td>71.23</td>
<td>79.73</td>
<td>79.75</td>
<td>8.52</td>
</tr>
<tr>
<td>&lt;15ms</td>
<td>83.14</td>
<td>83.11</td>
<td>88.14</td>
<td>88.16</td>
<td>5.05</td>
</tr>
<tr>
<td>&lt;20ms</td>
<td>88.94</td>
<td>88.97</td>
<td>92.09</td>
<td>92.11</td>
<td>3.14</td>
</tr>
</tbody>
</table>

The line with diamonds and the line with rectangles represent the FER results of the training (inside test) and testing sets, respectively. We observe that the ML-trained acoustic models yield FER of 10.31% and 11.77%, respectively, for the training and testing sets. In contrast, with 10 iterations, the MBE-trained acoustic models yield FER of 6.88% and 9.25%, respectively. The MBE discriminative training approach achieves a relative FER reduction of 33.27% on the training set and 21.41% on the testing set. The results clearly show that the MBE discriminative training approach performs very well and can enhance the performance of the acoustic models initially trained by using the ML criterion.

In our experiments, the baseline achieved an accuracy of 88.7% (under MSM configuration) for phonetic segmentation obtained by the MBE-trained models is very useful for subsequent manual verification or further boundary refinement using other techniques. The MBE training method is not difficult to implement, in particular some discriminative training tools, such as MPE, have been included in HTK.

In addition to applying the MBE criterion to the training of acoustic models, we have applied it to discriminative feature training. The preliminary experiment results indicate that feature-based MBE training is more effective than model-based MBE training. The segmentation accuracy could be improved by integrating the feature-based and model-based MBE training procedures. Moreover, a new decoding algorithm based on the minimum boundary error criterion is also under development. It is hoped that phone boundaries can be located more accurately by running a second pass search using the minimum boundary error criterion on the lattice generated by a first pass conventional search. In our current implementation, the phone boundary error function, defined in Eq.(5), is calculated in the time frame unit for efficiency. However, more accurate segmentation may be achieved by calculating boundary errors in actual time sample marks.

5. Acknowledgements
This work was funded by the National Science Council of the Republic of China under Grant: NSC94-2213-E-001-021.

6. References