A Minimum Boundary Error Framework for Automatic Phonetic Segmentation

Jen-Wei Kuo and Hsin-Min Wang

Institute of Information Science, Academia Sinica, Taipei, Taiwan
{rogerkuo, whm}@iis.sinica.edu.tw

Abstract. This paper presents a novel framework for HMM-based automatic phonetic segmentation that improves the accuracy of placing phone boundaries. In the framework, both training and segmentation approaches are proposed according to the minimum boundary error (MBE) criterion, which tries to minimize the expected boundary errors over a set of possible phonetic alignments. This framework is inspired by the recently proposed minimum phone error (MPE) training approach and the minimum Bayes risk decoding algorithm for automatic speech recognition. To evaluate the proposed MBE framework, we conduct automatic phonetic segmentation experiments on the TIMIT acoustic-phonetic continuous speech corpus. MBE segmentation with MBE-trained models can identify 80.53% of human-labeled phone boundaries within a tolerance of 10 ms, compared to 71.10% identified by conventional ML segmentation with ML-trained models. Moreover, by using the MBE framework, only 7.15% of automatically labeled phone boundaries have errors larger than 20 ms.

Keywords: automatic phonetic segmentation, minimum boundary error, discriminative training, minimum Bayes risk.

1 Introduction

Many areas of speech technology exploit automatic learning methodologies that rely on large well-labeled corpora. Phoneme level transcription is especially important for fundamental speech research. In recent years, increased attention has been paid to data-driven, concatenation-based TTS synthesis because its output is more natural and has a high degree of 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, the speech synthesis has relied on manually segmented corpora; however, such corpora are extremely hard to obtain, since labeling by hand is time consuming and costly. In speech recognition tasks, though the use of Hidden Markov Models (HMMs) has made finding precise phonetic boundaries unnecessary, it is believed that speech recognition would benefit from more precise segmentation in training and recognition.

To reduce the manual effort and accelerate 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 two-stage approaches [4].

The most popular 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 generally capture the acoustic properties of phones; however, small errors are inevitable because HMMs are not sensitive enough to detect changes between adjacent phones [4]. To improve the discriminability of HMMs for automatic phonetic segmentation, we proposed using a discriminative criterion, called the minimum boundary error (MBE), for model training in our previous work [5]. In this paper, the MBE criterion is extended to the segmentation stage, i.e., we propose an MBE forced alignment to replace the conventional maximum likelihood (ML) forced alignment. The superiority of the MBE framework over the conventional ML framework for automatic phonetic segmentation is verified by experiments conducted on the TIMIT acoustic-phonetic continuous speech corpus.

The remainder of this paper is organized as follows. Section 2 reviews the methodology of the MBE discriminative training approach. In Section 3, we present the proposed MBE segmentation approach and discuss its relation to the minimum Bayes risk (MBR) criterion. The experiment results are detailed in Section 4. Finally, in Section 5, we present our conclusions and suggest some future research directions.

2 Minimum Boundary Error Training

Let $\mathbf{O}=\{O^1,..,O^R\}$ be a set of training observation sequences. The objective function for MBE training can then be defined as:

$$F_{MBE} = \sum_{r=1}^{R} \sum_{S_i^r \in \mathbf{\Phi}^r} P(S_i^r \mid O^r) ER(S_i^r, S_c^r) , \tag{1}$$

where $\mathbf{\Phi}^r$ is a set of possible phonetic alignments for the training observation utterance O^r; S_i^r is one of the hypothesized alignments in $\mathbf{\Phi}^r$; $P(S_i^r \mid O^r)$ is the posterior probability of alignment S_i^r, given the training observation sequence O^r; and $ER(S_i^r, S_c^r)$ denotes the "boundary error" of S_i^r compared with the manually labeled phonetic alignment S_c^r. For each training observation sequence O^r, F_{MBE} gives the weighted average boundary error of all hypothesized alignments. For simplicity, we assume the prior probability of alignment S_i^r is uniformly distributed, and the likelihood $p(O^r \mid S_i^r)$ of alignment S_i^r is governed by the acoustic model parameter set Λ. Therefore, Eq.(1) can be rewritten as:

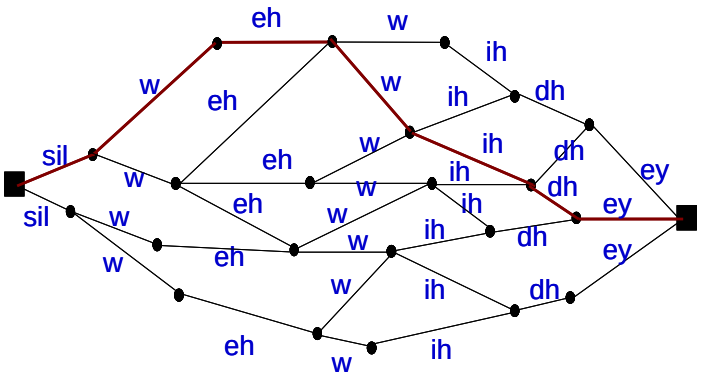

Fig. 1. An illustration of the phonetic lattice for the speech utterance: "Where were they?" The lattice can be generated by performing a beam search using some pruning techniques.

$$F_{MBE} = \sum_{r=1}^{R} \sum_{S_i^r \in \mathbf{\Phi}^r} \frac{p_\Lambda(O^r \mid S_i^r)^\alpha}{\sum_{S_k^r \in \mathbf{\Phi}^r} p_\Lambda(O^r \mid S_k^r)^\alpha} ER(S_i^r, S_c^r), \tag{2}$$

where α is a scaling factor that prevents the denominator $\sum_{S_k^r \in \mathbf{\Phi}^r} p_\Lambda(O^r \mid S_k^r)$ being dominated by only a few alignments. Accordingly, the optimal parameter set Λ^* can be estimated by minimizing the objective function defined in Eq.(2) as follows:

$$\Lambda^* = \arg\min_{\Lambda} \sum_{r=1}^{R} \sum_{S_i^r \in \mathbf{\Phi}^r} \frac{p_\Lambda(O^r \mid S_i^r)^\alpha}{\sum_{S_k^r \in \mathbf{\Phi}^r} p_\Lambda(O^r \mid S_k^r)^\alpha} ER(S_i^r, S_c^r). \tag{3}$$

The boundary error $ER(S_i^r, S_c^r)$ of the hypothesized alignment S_i^r can be calculated as the sum of the boundary errors of the individual phones in S_i^r, i.e.,

$$ER(S_i^r, S_c^r) = \sum_{n=1}^{N^r} er(q_n^i, q_n^c), \tag{4}$$

where N^r is the number of total phones in O^r; q_n^i and q_n^c are the *n*-th phone in S_i^r and S_c^r, respectively; and $er(\cdot)$ is a phone boundary error function defined as,

$$er(q_n^i, q_n^c) = 0.5 \times (\left| s_n^i - s_n^c \right| + \left| e_n^i - e_n^c \right|), \tag{5}$$

where s_n^i and e_n^i are the hypothesized start time and end time of phone q_n^i, respectively; and s_n^c and e_n^c correspond to the manually labeled start time and end

time, respectively. Since $\mathbf{\Phi}^r$ contains a large 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, such as an *N*-best list [6] or a lattice (or graph) [7], 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 Fig. 1, a phonetic lattice is more effective because it only stores alternative phone arcs on different segments of time marks and can easily generate a large number of distinct hypothesized phone alignments. Although it cannot be guaranteed that the phonetic alignments generated from a phonetic lattice will have higher probabilities than those not presented, we believe that the approximation will not affect the segmentation performance significantly. In this paper, we let $\mathbf{\Phi}^r_{\mathbf{Lat}}$ denote the set of possible phonetic alignments in the lattice for the training observation utterance O^r.

2.1 Objective Function Optimization and Update Formulae

Eq.(3) is a complex problem to solve, because there is no closed-form solution. In this paper, we adopt the Expectation Maximization (EM) algorithm to solve it. Since the EM algorithm maximizes the objective function, we reverse the sign of the objective function defined in Eq. (3) and re-formulate the optimization problem as,

$$\Lambda^* = \arg\max_{\Lambda} -\sum_{r=1}^{R}\sum_{S_i^r\in\mathbf{\Phi}^r}\frac{p_\Lambda(O^r \mid S_i^r)^\alpha}{\sum_{S_k^r\in\mathbf{\Phi}^r} p_\Lambda(O^r \mid S_k^r)^\alpha} ER(S_i^r, S_c^r)\,. \tag{6}$$

However, the EM algorithm can not be applied directly, because the objective function comprises rational functions [8]. The extended EM algorithm, which utilizes a weak-sense auxiliary function [9] and has been applied in the minimum phone error (MPE) discriminative training approach [10] for ASR, can be adapted to solve Eq.(6). The re-estimation formulae for the mean vector μ_m and the diagonal covariance matrix Σ_m of a given Gaussian mixture m thus derived can be expressed, respectively, as:

$$\mu_m = \frac{\theta_m^{MBE}(O) + D_m\overline{\mu}_m}{\gamma_m^{MBE} + D_m}\,, \tag{7}$$

and

$$\Sigma_m = \frac{\theta_m^{MBE}(O^2) + D_m\left[\overline{\Sigma}_m + \overline{\mu}_m\overline{\mu}_m^T\right]}{\gamma_m^{MBE} + D_m} - \mu_m\mu_m^T\,. \tag{8}$$

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 $\theta_m^{MBE}(O)$, $\theta_m^{MBE}(O^2)$, and γ_m^{MBE} are statistics defined, respectively, as:

$$\theta_m^{MBE}(O) = \sum_r \sum_{q \in \mathbf{\Phi}_{\mathbf{Lat}}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MBE} \gamma_{qm}^r(t) o_r(t) , \tag{9}$$

$$\theta_m^{MBE}(O^2) = \sum_r \sum_{q \in \mathbf{\Phi}_{\mathbf{Lat}}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MBE} \gamma_{qm}^r(t) o_r(t) o_r(t)^T , \tag{10}$$

and

$$\gamma_m^{MBE} = \sum_r \sum_{q \in \mathbf{\Phi}_{\mathbf{Lat}}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MBE} \gamma_{qm}^r(t) . \tag{11}$$

In Eqs. (9), (10), and (11), $\gamma_{qm}^r(t)$ is the occupation probability of mixture *m* on *q*, $o_r(t)$ is the observation vector at time *t*, and $\gamma_q^{r\,MBE}$ is computed by

$$\gamma_q^{r\,MBE} = \gamma_q^r \left(\eta_{avg}^r - \eta_q^r \right), \tag{12}$$

where γ_q^r is the occupation probability of phone arc *q*, also referred to as its posterior probability; η_{avg}^r is the weighted average boundary error of all the hypothesized alignments in the lattice; and η_q^r is the weighted average boundary error of 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 η_{avg}^r equals η_q^r, phone arc *q* makes no contribution to MBE training. However, when η_{avg}^r is larger than η_q^r, i.e., phone arc *q* generates fewer errors than the average, then *q* makes a positive contribution. Conversely, if η_{avg}^r is smaller than η_q^r, *q* makes a negative contribution. The discriminative ability of the MBE training approach is thus demonstrated. γ_q^r, η_{avg}^r, and η_q^r are computed by

$$\gamma_q^r = \frac{\sum_{S_i \in \mathbf{\Phi}_{\mathbf{Lat}}^r, q \in S_i} p_{\bar{\Lambda}}(O_r \mid S_i)^\alpha}{\sum_{S_k \in \mathbf{\Phi}_{\mathbf{Lat}}^r} p_{\bar{\Lambda}}(O_r \mid S_k)^\alpha} , \tag{13}$$

$$\eta_{avg}^{r} = \frac{\sum_{S_i \in \Phi_{\mathbf{Lat}}^{r}} p_{\bar{\Lambda}}(O_r \mid S_i)^{\alpha} ER(S_i)}{\sum_{S_k \in \Phi_{\mathbf{Lat}}^{r}} p_{\bar{\Lambda}}(O_r \mid S_k)^{\alpha}}, \tag{14}$$

and

$$\eta_{q}^{r} = \frac{\sum_{S_i \in \Phi_{\mathbf{Lat}}^{r}, q \in S_i} p_{\bar{\Lambda}}(O_r \mid S_i)^{\alpha} ER(S_i)}{\sum_{S_k \in \Phi_{\mathbf{Lat}}^{r}, q \in S_k} p_{\bar{\Lambda}}(O_r \mid S_k)^{\alpha}}, \tag{15}$$

respectively, where $\bar{\Lambda}$ 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 [10] 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 μ_m and the diagonal covariance matrix Σ_m thus become:

$$\mu_m = \frac{\theta_m^{MBE}(O) + D_m \bar{\mu}_m + \frac{\tau_m}{\gamma_m^{ML}} \theta_m^{ML}(O)}{\gamma_m^{MBE} + D_m + \tau_m}, \tag{16}$$

and

$$\Sigma_m = \frac{\theta_m^{MBE} + D_m \left[\bar{\Sigma}_m + \bar{\mu}_m \bar{\mu}_m^{T}\right] + \frac{\tau_m}{\gamma_m^{ML}} \theta_m^{ML}(O^2)}{\gamma_m^{MBE} + D_m + \tau_m} - \mu_m \mu_m^{T}, \tag{17}$$

respectively, where τ_m is also a per-mixture level control constant; and γ_m^{ML}, $\theta_m^{ML}(O)$, and $\theta_m^{ML}(O^2)$ are computed by

$$\gamma_m^{ML} = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_m^{r\,ML}(t), \tag{18}$$

$$\theta_m^{ML}(O) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_m^{r\,ML}(t) o_r(t), \tag{19}$$

and

$$\theta_m^{ML}(O^2) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_m^{r\,ML}(t) o_r(t) o_r(t)^T , \tag{20}$$

respectively. In Eqs. (18), (19), and (20), T_r is the frame number of O_r, and $\gamma_m^{r\,ML}(t)$ is the maximum likelihood occupation probability of the Gaussian mixture m.

3 Minimum Boundary Error Segmentation

The proposed MBE forced alignment approach is a promising realization of the *Minimum Bayes-Risk* (MBR) classifier for the automatic phonetic segmentation task. The latter can be considered as taking an action, $\alpha_S(O)$, to identify a certain alignment, S, from all the various phonetic alignments of a given utterance O. Let function $L(S, S_c)$ be the loss incurred when the action $\alpha_S(O)$ is taken, given that the true (or reference) alignment is S_c. During the classification stage, we do not know the true alignment in advance, i.e., any arbitrary alignment S_j could be true. Suppose the distribution $P(S_j \mid O)$ is known, then the conditional risk of taking the action $\alpha_S(O)$ is given by:

$$R(\alpha_S \mid O) = \sum_{S_j} L(S, S_j) P(S_j \mid O) . \tag{21}$$

The MBR classifier is designed to select the action whose conditional risk, $R(\alpha_S \mid O)$, is minimal, i.e., the best alignment based on the MBR criterion can be found by

$$S^* = \arg\min_S R(\alpha_S \mid O) = \arg\min_S \sum_{S_j \in \mathbf{\Phi}} L(S, S_j) P(S_j \mid O) . \tag{22}$$

When the symmetrical zero-one function,

$$L(S, S_j) = \begin{cases} 0, & S = S_j \\ 1, & S \neq S_j \end{cases} , \tag{23}$$

is selected as the loss function, and it is assumed that the prior probability of alignment S_j is uniformly distributed, the MBR classifier is equivalent to the conventional forced-alignment method, which picks the alignment with the maximal likelihood, i.e.,

$$
\begin{aligned}
S^* &= \arg\min_S \sum_{S_j \in \mathbf{\Phi}} L(S, S_j) P(S_j \mid O) \\
&= \arg\min_S \sum_{S_j \in \mathbf{\Phi}, S_j \neq S} P(S_j \mid O) \\
&= \arg\min_S (1 - P(S \mid O)) \\
&= \arg\max_S P(O \mid S)
\end{aligned}
\quad . \tag{24}
$$

It is clear from Eq. (23) that the zero-one loss function assigns no loss when $S = S_j$, but assigns a uniform loss of one to the alignments $S \neq S_j$ no matter how different they are from S_j. Thus, such a loss function causes all incorrectly hypothesized alignments to be regarded as having the same segmentation risk, which is obviously inconsistent with our preference for alignments with fewer errors in an automatic segmentation task.

In our approach, the loss function is replaced by the boundary error function, defined in Eq.(4), to match the goal of minimizing the boundary error. Consequently, the MBR forced alignment approach becomes the MBE forced alignment approach, defined as:

$$
\begin{aligned}
S^* &= \arg\min_S \sum_{S_j \in \mathbf{\Phi}} ER(S, S_j) P(S_j \mid O) \\
&= \arg\min_S \sum_{S_j \in \mathbf{\Phi}} \sum_{n=1}^{N} er(q_n, q_n^j) P(S_j \mid O)
\end{aligned}
\quad , \tag{25}
$$

where N is the number of phones in utterance O; and q_n and q_n^j are the n-th phone in the alignments S and S_j, respectively.

To simplify the implementation, we restrict the hypothesized space $\mathbf{\Phi}$ to $\mathbf{\Phi}_{\mathbf{Lat}}$, the set of alignments constructed from the phone lattice shown in Fig. 1, which can be generated by a conventional beam search. Accordingly, Eq. (25) can be re-formulated as:

$$
\begin{aligned}
S^* &= \arg\min_S \sum_{S_j \in \mathbf{\Phi}_{\mathbf{Lat}}} \sum_{n=1}^{N} P(S_j \mid O) er(q_n, q_n^j) \\
&= \arg\min_S \sum_{n=1}^{N} \sum_{S_j \in \mathbf{\Phi}_{\mathbf{Lat}}} P(S_j \mid O) er(q_n, q_n^j)
\end{aligned}
\quad . \tag{26}
$$

Let the ***cut*** $\mathbf{C}_n$ be the set of phone arcs of the n-th phone in the utterance. For example, in Fig. 1, there are four phone arcs for the second phone, "w", in $\mathbf{C}_2$ and six phone arcs for the third phone, "eh", in $\mathbf{C}_3$. From the figure, it is obvious that

each alignment in $\mathbf{\Phi_{Lat}}$ will pass a single phone arc in each ***cut*** $\mathbf{C}_n$, n=1,2,…,N. According to this observation, Eq. (26) can be rewritten as:

$$S^* = \arg\min_{S} \sum_{n=1}^{N} \sum_{q_{n,m} \in \mathbf{C}_n} \sum_{\{S_j \in \mathbf{\Phi_{Lat}} | q_{n,m} \in S_j\}} P(S_j \mid O) er(q_n, q_{n,m}) , \tag{27}$$

where $q_{n,m}$ is the m-th phone arc in $\mathbf{C}_n$. Because $\sum_{\{S_j \in \mathbf{\Phi_{Lat}} | q_{n,m} \in S_j\}} P(S_j \mid O)$ in Eq. (27) is equivalent to the posterior probability of $q_{n,m}$ given the utterance O, denoted as $\gamma_{q_{n,m}}$ hereafter, the probability can be easily calculated by applying a forward-backward algorithm to the lattice. As a result, Eq. (27) can be rewritten as:

$$S^* = \arg\min_{S} \sum_{n=1}^{N} \sum_{q_{n,m} \in \mathbf{C}_n} \gamma_{q_{n,m}} er(q_n, q_{n,m}) . \tag{28}$$

In this way, MBE forced alignment can be efficiently conducted on the phone lattice by performing Viterbi search.

4 Experiments

4.1 Experiment Setup

TIMIT (The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus) [11], a well-known read speech corpus with manual acoustic phonetic labeling, has been widely used to evaluate automatic speech recognition and phonetic segmentation techniques. TIMIT contains a total of 6,300 sentences spoken by 630 speakers from eight major dialect regions in the United States; each speaker utters 10 sentences. 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 test 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, plus 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 test speech utterances.

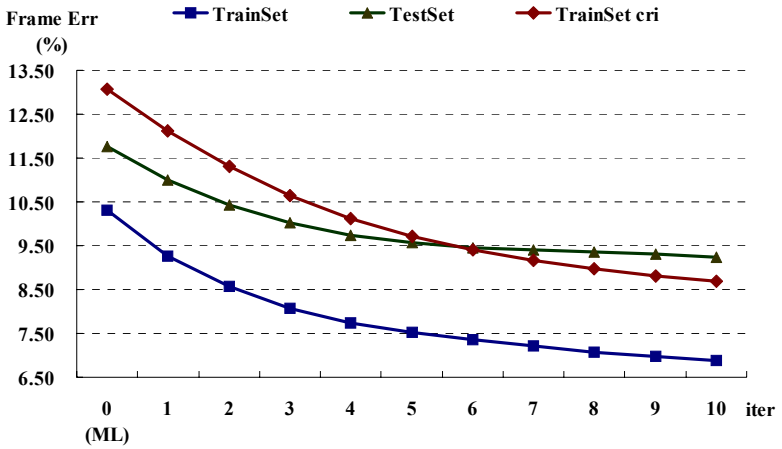

Fig. 2. The phonetic segmentation results (FER) for the models trained according to ML and MBE criteria, respectively.

4.2 Experiment Results

The acoustic models were first trained on the training utterances according to human-labeled phonetic transcriptions and boundaries by the Baum-Welch algorithm using the ML criterion. Then, the MBE discriminative training approach was applied to further manipulate the models. The scaling factor α in Eq.(2) was empirically set to 0.1 and the I-smoothing control constant τ_m in Eqs.(16) and (17) was set to 20 for all mixtures. The results are shown in Fig. 2. In the figure, the line with triangles indicates the expected FER (frame error rate) calculated at each iteration of the training process. Clearly, the descending trend satisfies the training criterion. The line with diamonds and the line with rectangles represent the FER results of the training (inside test) and test sets, respectively. We observe that the ML-trained acoustic models (at the 0th iteration) yield an FER of 10.31% and 11.77% for the training set and test set respectively. In contrast, after 10 iterations, the MBE-trained acoustic models yield an 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 test set. The results clearly demonstrate that the MBE discriminative training approach performs very well and can enhance the performance of the acoustic models initially trained by the ML criterion.

Table 1 shows the percentage of phone boundaries correctly placed within different tolerances with respect to their associated manually-labeled phone boundaries. The experiment was conducted on the test set. From rows 2 and 3 of Table 1, we observe that the MBE-trained models significantly outperform the ML-trained models. Clearly, the MBE training is particularly effective in correcting boundary errors in the proximity of manually labeled positions. Comparing the results in rows 2 and 4, we also observe that MBE segmentation outperforms ML segmentation, though the

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

Criterion		Mean Boundary Distance	%Correct marks (distance ≦ tolerance)					
Training	Segmentation		≦5ms	≦10ms	≦15ms	≦20ms	≦25ms	≦30ms
ML	ML	9.83 ms	46.69	71.10	83.14	88.94	92.32	94.52
ML+MBE	ML	7.82 ms	58.48	79.75	88.16	92.11	94.49	96.11
ML	MBE	8.95 ms	49.86	74.25	85.38	90.61	93.75	95.67
ML+MBE	MBE	7.49 ms	58.73	80.53	88.97	92.85	95.16	96.64
absolute improvement (ML+MBE,MBE) vs. (ML, ML)		2.34 ms	12.04	9.43	5.83	3.91	2.84	2.12

improvement is not as significant as that of the MBE-trained models over the ML-trained models. This is because, MBE segmentation, like conventional ML segmentation, is still deficient in the knowledge of true posterior distribution, even though the MBE criterion accords with the objective of minimizing boundary errors very well. The 5th row of Table 1 shows the results obtained when the complete MBE framework, including MBE training and MBE segmentation, was applied. We observe that these results are superior to those achieved when either the MBE training or the MBE segmentation was applied alone. The last row of Table 1 shows the absolute improvements achieved by the MBE framework over the conventional ML framework. The proposed MBE framework can identify 80.53% of human-labeled phone boundaries within a tolerance of 10 ms, compared to 71.10% identified by the conventional ML framework. Moreover, by using the MBE framework, only 7.15% of automatically labeled phone boundaries have errors larger than 20 ms.

5 Conclusions and Future Work

In this paper, we have explored the use of the minimum boundary error (MBE) criterion in the discriminative training of acoustic models as well as minimum risk segmentation for automatic phonetic segmentation. The underlying characteristics of the MBE training and segmentation framework have been investigated, and its superiority over conventional ML training and segmentation has been verified by experiments. Naturally, the more accurate phonetic segmentation obtained by the MBE framework is very useful for subsequent manual verification or further boundary refinement using other techniques. It is worth mentioning that the MBE training method is not difficult to implement; in particular, minimum phone error training has been included in HTK.

In HMM-based automatic phonetic segmentation and speech recognition tasks, duration control is an important issue that must be addressed. We tried to apply the

MBE criterion in duration model training, but there was no significant improvement found in our preliminary work. However, the issue warrants further study. On the other hand, well-labeled phonetic training corpora are very scarce. Therefore, the unsupervised MBE training approach is also under investigated. Moreover, 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. In addition, we are applying the MBE training and segmentation framework to facilitate the phonetic labeling of a subset of speech utterances in MATBN (Mandarin across Taiwan − Broadcast News) database [12].

Acknowledgment. This work was funded by the National Science Council, Taiwan, under Grant: NSC94-2213-E-001-021.

References

1. Malfrere, F., Dutiot, T.: High-quality speech synthesis for phonetic speech segmentation. Proc. Fifth Eurospeech (1997) 2631-2634
2. van Santen, J., Sproat, R.: High accuracy automatic segmentation. Proc. Sixth Eurospeech (1999) 2809-2812
3. Brugnara, F., Falavigna, D., Omologo, M.: Automatic segmentation and labeling of speech based on Hidden Markov Models. Speech Communication, Vol. 12, Issue. 4 (1993) 357-370
4. Torre Toledano, D., Rodriguez Crespo, M. A., Escalada Sardina, J. G.: Try to mimic human segmentation of speech using HMM and fuzzy logic post-correction rules. Proc. Third ESCA/COCOSDA International Workshop on Speech Synthesis (1998) 1263-1266
5. Kuo, J.-W., Wang, H.-M.: Minimum Boundary Error Training for Automatic Phonetic Segmentation. Proc. Interspeech – ICSLP (2006)
6. Schwartz, R., Chow, Y.-L.: The N-best algorithms: an efficient and exact procedure for finding the N most likely sentence hypotheses. Proc. ICASSP, Vol. 1(1990) 81-84
7. Ortmanns, S., Ney, H., Aubert, X.: A word graph algorithm for large vocabulary continuous speech recognition. Computer Speech and Language, Vol. 11 (1997) 43-72
8. Gopalakrishnan, P., Kanevsky, D., Nádas, A., Nahamoo, D.: An inequality for rational functions with applications to some statistical estimation problems. IEEE Trans. Information Theory, Vol. 37 (1991) 107-113
9. Povey, D.: Discriminative Training for Large Vocabulary Speech Recognition. Ph.D. Dissertation, Peterhouse, University of Cambridge, July 2004
10. Povey, D., Woodland, P. C.: Minimum phone error and I-smoothing for improved discriminative training. Proc. ICASSP, Vol. 1 (2002) 105-108
11. Lamel, L., Kasel, R., Seneff, S.: Speech database development: design and analysis of the acoustic-phonetic corpus. Proc. DARPA Speech Recognition Workshop (1986) 100-109
12. Wang, H.-M., Chen, B., Kuo, J.-W., Cheng, S.-S.: MATBN: A Mandarin Chinese Broadcast News Corpus. International Journal of Computational Linguistics & Chinese Language Processing, Vol. 10, No. 2, June (2005) 219-236