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 two stage approaches [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 $O=\{O_1,..,O_R\}$, 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:

$$F_{MBE} = \sum_{r=1}^{R} \sum_{S_i \in \mathbf{\Phi}_r} P(S_i \mid O_r) ER(S_i), \quad (1)$$

where $\mathbf{\Phi}_r$ is a set of various possible phonetic alignments for the training observation utterance O_r; S_i is one of the hypothesized alignments in $\mathbf{\Phi}_r$; $P(S_i \mid O_r)$ is the posterior probability of alignment S_i, given the training observation sequence O_r; and $ER(S_i)$ denotes the "boundary error" of S_i compared with the manually labeled phonetic alignment in the canonical transcription. 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 is uniformly

distributed, and the likelihood $P(O_r \mid S_i)$ of alignment S_i is governed by the acoustic model parameter set Λ. Therefore, Eq.(1) can be rewritten as:

$$F_{MBE} = \sum_{r=1}^{R} \sum_{S_i \in \mathbf{\Phi}_r} \frac{P_\Lambda(O_r \mid S_i)^\alpha}{\sum_{S_k \in \mathbf{\Phi}_r} P_\Lambda(O_r \mid S_k)^\alpha} ER(S_i), \quad (2)$$

where α is a scaling factor that prevents the denominator $\sum_{S_k \in \mathbf{\Phi}_r} P_\Lambda(O_r \mid S_k)$ being dominated by only a few alignments. If α is set to zero, all the hypotheses are equally weighted. Accordingly, the optimal parameter set Λ^* can be estimated by minimizing the objective function defined in Eq.(2), i.e.,

$$\Lambda^* = \arg\min_{\Lambda} \sum_{r=1}^{R} \sum_{S_i \in \mathbf{\Phi}_r} \frac{P_\Lambda(O_r \mid S_i)^\alpha}{\sum_{S_k \in \mathbf{\Phi}_r} P_\Lambda(O_r \mid S_k)^\alpha} ER(S_i). \quad (3)$$

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

$$ER(S_i) = \sum_{q \in S_i} er(q), \quad (4)$$

where q is a phone involved in S_i; $er(\cdot)$ is a phone boundary error function defined as,

$$er(q) = 0.5 \times (\left|s_q - s'_q\right| + \left|e_q - e'_q\right|), \quad (5)$$

where s_q and e_q are, respectively, the hypothesized start time and end time of phone q; and s'_q and e'_q correspond to the manually labeled start time and end time, respectively. Since $\mathbf{\Phi}_r$ 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, such as 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 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 all 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.

2.1 Objective function optimization and update formulae

Eq.(3) is a complex problem to solve because there is no closed-form solution. Even so, some iterative techniques, such as the Expectation Maximization (EM) algorithm, can be applied to solve it. Since the EM algorithm maximizes the objective function, we reverse the sign of our objective function and re-formulate the optimization problem as,

$$\Lambda^* = \arg\max_{\Lambda} - \sum_{r=1}^{R} \sum_{S_i \in \mathbf{\Phi}_r} \frac{P_\Lambda(O_r \mid S_i)^\alpha}{\sum_{S_k \in \mathbf{\Phi}_r} P_\Lambda(O_r \mid S_k)^\alpha} ER(S_i). \quad (6)$$

However, the EM algorithm can not be applied directly because the objective function comprises rational functions [7]. The extended EM algorithm, which utilizes a weak-sense

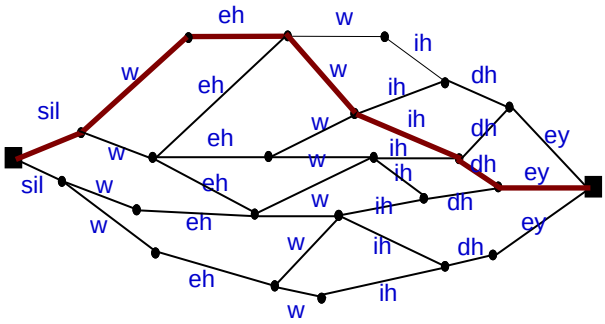

Figure 1: *An illustration of the phonetic lattice for the speech utterance "where were they?". The lattice can be generated by performing a beam search using some pruning techniques.*

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 μ_m and the diagonal covariance matrix Σ_m of a given Gaussian mixture m thus derived can be expressed, respectively, as:

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

and

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

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

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

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

and

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

In Eqs. (9), (10), and (11), $\gamma_{qm}^r(t)$ is the occupation probability for mixture m on q, $o_r(t)$ is the observation vector at time t, and $\mathbf{\Phi}_r^{lat}$ represents the lattice for sentence O_r. $\gamma_q^{r\,MBE}$ is computed by

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

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 shown. γ_q^r, η_{avg}^r, and η_q^r are computed by

$$\gamma_q^r = \frac{\sum_{S_i \in \Phi_r^{lat}, q \in S_i} p_{\overline{\Lambda}}(O_r \mid S_i)^{\alpha}}{\sum_{S_k \in \Phi_r^{lat}} p_{\overline{\Lambda}}(O_r \mid S_k)^{\alpha}}, \tag{13}$$

$$\eta_{avg}^r = \frac{\sum_{S_i \in \Phi_r^{lat}} p_{\overline{\Lambda}}(O_r \mid S_i)^{\alpha} ER(S_i)}{\sum_{S_k \in \Phi_r^{lat}} p_{\overline{\Lambda}}(O_r \mid S_k)^{\alpha}}, \tag{14}$$

and

$$\eta_q^r = \frac{\sum_{S_i \in \Phi_r^{lat}, q \in S_i} p_{\overline{\Lambda}}(O_r \mid S_i)^{\alpha} ER(S_i)}{\sum_{S_k \in \Phi_r^{lat}, q \in S_k} p_{\overline{\Lambda}}(O_r \mid S_k)^{\alpha}}, \tag{15}$$

respectively, where $\overline{\Lambda}$ 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 μ_m and the diagonal covariance matrix Σ_m thus become:

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

and

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

respectively, where τ_m is also a per-mixture level control constant, and

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

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

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

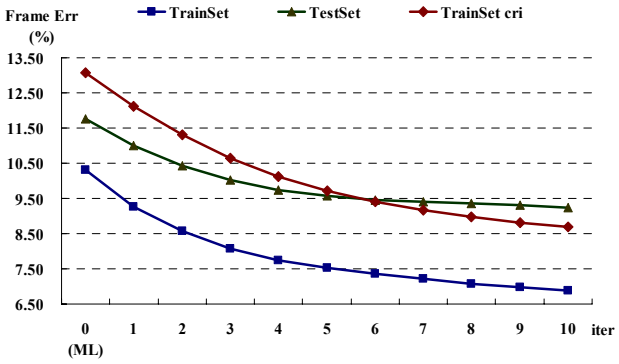

Figure 2: *The phonetic segmentation results (FER) for the models trained by MBE with I-smoothing applied.*

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 Gaussian mixture *m*.

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 α in Eq.(2) was 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 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.

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

	Mean Boundary Distance	%Correct marks (error < tolerance)			
		<5ms	<10ms	<15ms	<20ms
ML_{10}	9.83 ms	46.69	71.10	83.14	88.94
ML_{20}	9.78 ms	46.95	71.23	83.11	88.97
ML_{10} + MBE_{10}	7.83 ms	58.35	79.73	88.14	92.09
ML_{10} + MBE_{10} + I-smoothing	7.82 ms	58.48	79.75	88.16	92.11
absolute improvement	1.96 ms	11.53	8.52	5.05	3.14

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.

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 testing set. We observe that the MBE-trained models with 10 iterations (ML_{10}+MBE_{10}) outperform their seed models, i.e., the ML-trained models with 10 iterations (ML_{10}), and the ML-trained models with 20 iterations (ML_{20}). We also observe that the I-smoothing technique can only slightly improve the performance. The last row of Table 1 shows the absolute improvements of the best results (ML_{10} + MBE_{10}+ I-smoothing) compared to the results of ML_{20}. It is clear that the MBE training is particularly effective in correcting boundary errors in the proximity of manually labeled positions. In [3], Brugnara *et al.* presented an excellent HMM-based phonetic segmentation system, which achieved an accuracy of 88.7% (under MSM configuration) within a tolerance of 20ms. In our experiments, the baseline HMM-based system yields an accuracy of 88.97% (ML_{20}), while the MBE training improves the accuracy to 92.11%.

4. Conclusions and future work

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

[1] F. Malfrere and T. Dutiot, "High-quality speech synthesis for phonetic speech segmentation," in *Proc. Eurospeech'97*, pp.2631-2634.

[2] J. van Santen and R. Sproat, "High accuracy automatic segmentation," in *Proc. Eurospeech'99*, pp.2809-2812.

[3] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, Vol. 12, Issue. 4, pp.357-370, 1993.

[4] D. Torre Toledano, M. A. Rodriguez Crespo, and J. G. Escalada Sardina, "Try to mimic human segmentation of speech using HMM and fuzzy logic post-correction rules," in *Proc. Third ESCA/COCOSDA International Workshop on Speech Synthesis*, 1998, pp.1263-1266.

[5] R. Schwartz and Y.-L. Chow, "The N-best algorithms: an efficient and exact procedure for finding the N most likely sentence hypotheses," in *Proc. ICASSP'90.*

[6] S. Ortmanns, H. Ney, and X. Aubert, "A word graph algorithm for large vocabulary continuous speech recognition," *Computer Speech and Language*, Vol. 11, pp.43-72, 1997.

[7] P. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Information Theory*, Vol. 37, pp.107-113, 1991.

[8] D. Povey, Discriminative Training for Large Vocabulary Speech Recognition. *Ph.D. Dissertation, Peterhouse, University of Cambridge*, July 2004.

[9] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP'02.*

[10] L. Lamel, R. Kasel, and S. Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," in *Proc. DARPA Speech Recognition Workshop*, 1986, pp.100-109.