CHAPTER 7

Selected Experimental Results

In this chapter, we present experimental results on several automatic speech recognition (ASR) tasks. We evaluate the growth transformation (GT)-based minimum classification error (MCE) training method on both small-vocabulary, well-controlled benchmark tests such as TIDIGITS, and on large-vocabulary, real-world speech recognition tasks such as commercial telephony large-vocabulary ASR (LVASR) applications. We show that the GT-based discriminative training gives superior performance over the conventional maximum likelihood (ML)-based training method.

7.1 EXPERIMENTAL RESULTS ON SMALL ASR TASKS TIDIGITS

TIDIGITS is a speaker-independent connected-digit task. Each utterance in this corpus has an unknown length with a maximum of seven digits. The training set includes 8623 utterances and testing set includes 8700 utterances. In our experiment, a word-based hidden Markov model (HMM) is built for each of the 10 digits from ZERO to NINE, plus word OH. The number of states of each HMM ranges from 9 to 15, depending on the average duration of each word, and each state has an average of six Gaussian mixture components. The speech feature vector is computed from audio signal analysis, which gives 12 Mel frequency cepstral coefficients (MFCCs) and the audio energy, plus their first-order and second-order temporal differences.

GT-MCE training is performed with initialization from the ML-trained models. In this experiment, 1-best MCE training is used. The constant factor E as discussed in Section 5.3 is set to 1.0, and the sloping factor α is set to 0.01.

The experimental results are shown in Table 7.1. The baseline HMMs are trained with the ML criterion. The ML model gives a word error rate (WER) of 0.30% on testing data. With GT/extended Baum–Welch (EBW)-based MCE training, after three iterations, the algorithm convergence is reached and the WER is reduced to 0.23%. As a comparison, a conventional generalized probabilistic descent (GPD)-based MCE is also implemented for this task. As shown in Table 7.1, the best GPD MCE result is with a WER of 0.24%, which is obtained after 12 iterations over the full training data set (i.e., 12 epochs). The results of this small-task experiment show that the new GT-based MCE learning method is slightly better than the conventional GPD-based MCE,

TABLE 7.1: Comparative recognition-accuracy performance (measured by WER — the lower the better) of the new and traditional MCE training methods, as well as the ML method

	ML	GPD MCE	GT/EBW MCE
WER	0.30%	0.24%	0.23%
WER reduction	–	20.0%	23.3%

and it gives significantly improved efficiency in the training by providing much faster algorithm convergence.

Figure 7.1 gives a detailed analysis of the GT-based MCE learning. Figure 7.1a shows that the value of the objective function of MCE training decreases monotonically for each new iteration. Note that the loss function of the training set is the one defined by (3.18). Figure 7.1b shows that the WER on the test set is reduced significantly also after a few GT iterations. It is noteworthy that given different settings of the global constant E as discussed in Chapter 5.3, the model updating speed can be controlled. For example, a smaller E leads to faster training speed. However, the models trained this way may suffer from overtraining and hence cause unstable performance on the test set.

7.2 TELEPHONY LVASR APPLICATIONS

We further evaluated the GT-based discriminative training method on large vocabulary telephony speech recognition tasks. The Microsoft large-scale telephony speech databases are used to build a large-vocabulary telephony ASR system. The entire training set, which is collected through various channels including close-talk telephones, far-field microphones, and cell phones, consists of 26 separate corpuses, 2.7 million utterances, and a total of 2000 hours of speech data. To improve the robustness of the acoustic model, speech data are recorded under various conditions with different environmental noises and include both native English speakers and speakers with various foreign accents. The text prompts include common telephony-application style utterances and some dictation-style utterances from the *Wall Street Journal* database.

The test sets consist of several typical context-free grammar (CFG)-based commercial telephony ASR tasks. To evaluate the generalization ability of our approach, the test data are collected in a very different setup than the training set. The default global vocabulary size of the ASR system is 120K. However, different vocabularies are used in different test sets. Table 7.2 summarizes the test sets used in our experiments.

In this experiment, all data are sampled at a rate of 8K Hz. Phonetic decision trees are used for state tying and there are about 6000 tied states with an average of 16 Gaussian mixture components per state. The 52-dimensional raw acoustic feature vectors are composed of the normalized energy,

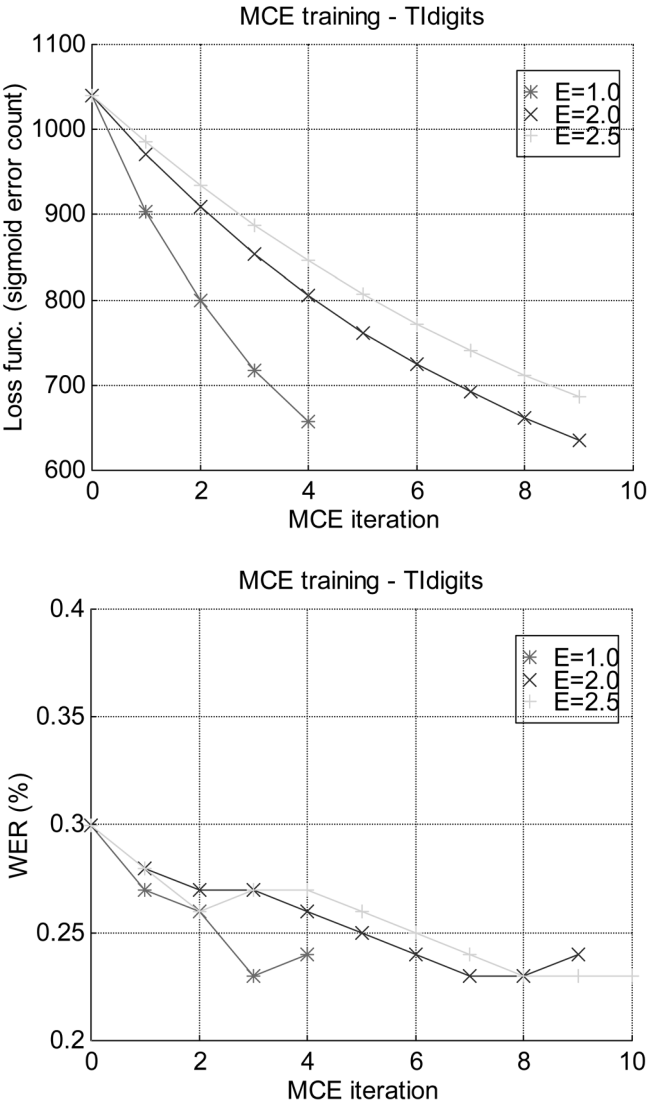

FIGURE 7.1: (a) GT-based MCE training. Number of iterations versus loss function of the training set given different settings of the global constant E that controls the model updating speed. (b) GT-based MCE training. Number of iterations versus WER on the test set given different settings of the global constant E that controls the model updating speed.

and 12 MFCCs and their first-, second-, and third-order time derivatives. The 52-dimensional raw features are further projected to form 36-dimensional feature vectors via heteroscedastic linear discriminant analysis transformation.

TABLE 7.2: Description of the test sets

NAME	VOCABULARY SIZE	WORD COUNT	DESCRIPTION
MSCT	70K	4356	General call center application.
STK	40K	12,851	Finance applications (stock transaction, etc.)
QSR	55K	5718	Name dialing application (note: pronunciations of most names are generated by letter-to-sound rules).

As with the TIDIGITS database, GT-MCE training is performed with initialization from the ML-trained models. The sloping facto α is set to 1/30 and the E factor is set to 1.0. In MCE training of HMMs, the training data are first decoded by a simple unigram weighted CFG containing all the words in the transcripts of training data and the competitors are then updated after each set of three iterations. All HMM parameters (except transition probabilities and Gaussian mixture weights) are updated. To prevent variance underflow, a dimension-dependent variance floor is set as 1/20 of the average variance over all Gaussian components in that dimension. The variance values

TABLE 7.3: Experimental results on three telephony ASR test sets

TEST SET		ML	MCE
MSCT	WER	12.413%	10.514%
	Abs. WERR Over ML	N/A	1.899%
	Rel. WERR Over ML	N/A	15.30%
STK	WER	7.993%	7.330%
	Abs. WERR Over ML	N/A	0.663%
	Rel. WERR Over ML	N/A	8.30%
QSR	WER	9.349%	8.464%
	Abs. WERR Over ML	N/A	0.885%
	Rel. WERR Over ML	N/A	9.47%
Average	WER	9.918%	8.769%
	Abs. WERR Over ML	N/A	1.149%
	Rel. WERR Over ML	N/A	11.58%

Q1

that are lower than the variance floor are set to the floor value. In the experiments, only six iterations (i.e., six epochs) are performed in MCE training.

Table 7.3 presents the WER on the three test sets. Compared with the ML baseline, the MCE training can reduce the WER by 11.58%. These results demonstrate that the MCE training approach has strong generalization ability. It can be effectively applied not only to small-scale tasks but also to large-scale ASR tasks.

Author Query Form

(Queries are to be answered by the Author)

He – Chapter 7

The following queries have arisen during the typesetting of your manuscript. Please answer these queries.

Query Marker	Query	Reply
Q1	Please provide a legend for the these abbreviations: MSCT, STK, QSR, N/A, Rel., Abs	

Thank you very much.