PRIOR KNOWLEDGE GUIDED MEL BASED MODEL SELECTION AND ADAPTATION FOR NONNATIVE SPEECH RECOGNITION

Xiaodong He and Yunxin Zhao

Dept. of Computer Engineering and Computer Science
University of Missouri, Columbia, MO 65211, USA

ABSTRACT

In this paper, an improved method of model complexity selection for nonnative speech recognition is proposed by using maximum *a posteriori* estimation of bias distributions. An algorithm is described for estimating the hyper-parameters of the prior distributions, and an automatic accent detection algorithm is also proposed for integration with dynamic model selection and adaptation. Experiments were performed on the WSJ1 task with American English speech, British accent speech, and mandarin Chinese accent speech. Results show that the use of prior knowledge of accents enabled reliable estimation of bias distributions in the case of very small amount of adaptation speech, or without adaptation speech. Recognition results show that the new approach is superior to the previous MEL method, especially when the adaptation data are extremely limited.

1. INTRODUCTION

English speech recognition systems are commonly trained from speech data of native English speakers. For certain tasks, these systems may work very well for native talkers, but the performance in general degrades drastically on speech with heavy foreign accents. It is difficult to train acoustic models for each foreign accent since the required vast amount of training data that covers different types and degrees of foreign accents do not yet exist.

Currently, improving recognition performance of nonnative speech is an active area of research [1] - [4]. One straightforward approach is to use speaker adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) [5] or Maximum a posteriori (MAP) estimation [6] to adapt speaker-independent models to the foreign-accent of a new speaker. It is well known that a much larger amount of adaptation speech data is needed from a foreign-accent speaker than a native speaker to achieve a comparable level of recognition accuracy [1]. Another approach is based on multilingual acoustic modeling, where the phone sets of several languages are mapped to a universal phone set and multilingual speech data are pooled to train the acoustic model [2]. So far, the multilingual approach has been limited to small tasks, and it was reported in [2] that compared with using acoustic model trained from native speech alone, multilingual acoustic model improved nonnative speech recognition but it degraded native speech recognition. Other techniques for nonnative speech recognition include nonnative speech based lexicon modeling, acoustic modeling, and decoding techniques [3] [4]. Such techniques required that the target foreign accent is known and substantial knowledge of foreign languages is built into ASR systems.

A model selection based speaker adaptation strategy for nonnative speakers has been proposed recently by the authors of the current paper [7] [8]. The method combines dynamic selection of model complexity with MLLR based model adaptation. Experimental results of [7] showed that between native speakers and nonnative speakers, the curves of model complexity vs. recognition performance were significantly different. Highly detailed acoustic models that produced the best recognition result for native speakers were worst for nonnative speakers, while intermediate levels of model complexity as determined from adaptation speech worked best for foreign accent talkers. In [8], model selection from using a small amount of adaptation speech was accomplished by a maximum expected likelihood (MEL) algorithm.

The MEL method consists of three steps. In the first step, an acoustic model based on phonetic decision trees (PDT) for triphone HMMs is trained from native English speech, where a Gaussian mixture density (GMD) is estimated for each node of a PDT, including tree internal nodes. In the second step, Viterbi alignment is performed on adaptation data and each feature vector is assigned to a dominant Gaussian component density (GC) of a terminal tree node, and for each GC of a terminal tree node that has adaptation data a bias is calculated between the data sample mean and the model mean. Within each tree node, the distribution parameters of the biases are estimated based on the assumption that the biases are Gaussian random variables, and the expected log-likelihood is then computed for the adaptation data. In the third step, the optimal tree cut, or model complexity, is determined to maximize the expected log-likelihood (EL) over tree cuts by using a bottom-up pruning method.

In the MEL method, the performance of model selection depends on the quality of the estimated bias distributions, and model selection could be unreliable when data are limited. As described in [8], a dynamic clustering scheme is deployed to group similar Gaussian components into an allophone cluster that corresponds to a tree node. In order to reliably estimation a bias distribution, a cluster is generated only when there are sufficient samples of bias accumulated in it. As the result, certain tree nodes would have bias distributions while others not, and a tree node without a bias distribution will then use the one from

This work is supported in part by National Institutes of Health under the grant NIH 1 R01 dc04340-01A2 and the National Science Foundation under the grant NSF EIA 9911095

its closest parent node. It is clear that when data is very sparse, only a few clusters might be generated, and each cluster may include GCs with very different properties. In the extreme case, all GCs could be grouped into a single cluster and share only one global bias distribution. This situation is undesirable since for each foreign accent, the data-model mismatch is highly dependent on phones or phone-classes, and sharing only a few, or a global, bias distributions over all phones would be to coarse to characterize a foreign accent.

If certain *a priori* knowledge of foreign accents could be utilized in estimating bias distributions, more reliable and detailed bias distributions could be obtained even when data is very limited. With prior knowledge, when the amount of data is small, parameter estimation is base more on prior knowledge than data to obtain better results. Moreover, given the prior knowledge of the mismatch condition at the level of phones, a bias distribution can be estimated for a phone unit that may not even have any data. In this way, bias distribution sharing can be made at a much finer level, and better performance of model selection is expected.

In this paper, a prior knowledge guided MEL approach, referred to as P-MEL, is proposed based on maximum *a posteriori* (MAP) estimation [9] of the bias distribution parameters. To facilitate choosing accent-specific knowledge for each speaker, an automatic accent detection algorithm is also developed. This new approach is evaluated on the WSJ task with speech data of American English speakers, British English speakers and mandarin Chinese accent English speakers. For each foreign accent, a small set of speech data was used to estimate the priors of the bias distributions. Recognition evaluation test was the 5000 words WSJ task. Experimental results verified that the P-MEL approach is superior to the MEL approach when the amount of adaptation data is very limited.

2. MAP ESTIMATION OF BIAS DISTRIBUTIONS

2.1. MAP based parameter estimation

Given a data set X, MAP estimation gives the optimal model by

$$\Lambda_{MAP} = \arg\max_{\Lambda} f(X \mid \Lambda) g(\Lambda), \qquad (1)$$

where the prior distribution $g(\Lambda)$ characterizes the knowledge about the model parameter set Λ. In general, when the size of X is small, $g(\Lambda)$ dominates (1) and the optimal Λ is based more on the prior knowledge; when the size of X is large, $f(X \mid \Lambda)$ dominates (1) and the optimal Λ is based more on the observed data. For each foreign accent, a set of prior distributions can be defined, with one distribution for one phone unit. The MAP estimate of Λ depends on assumption of $g(\Lambda)$ which is often taken as a conjugate prior distribution [9].

Assume for a Gaussian pdf $f(b|\Lambda)$ with $\Lambda = \{\mu, \theta\}$, where $\theta = 1/\sigma^2$ is the precision parameter. The joint conjugate prior $g(\mu,\theta)$ is a normal-gamma distribution, where the conditional distribution of μ given θ is a normal distribution with mean ν and variance $1/\tau\theta$, and the marginal distribution of θ is a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, i.e.,

$$g(\mu,\theta) = \frac{\sqrt{\tau\theta}}{\sqrt{2\pi}} \exp\left[-\frac{\tau\theta}{2}(\mu-\nu)^2\right] \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta) \qquad (2)$$

By setting $\alpha = \frac{\tau+1}{2}$, $\beta = \frac{\tau}{2}s^2$, the joint MAP estimate of μ and σ^2 given a set of sample biases $\{b_i\}$ is solved as [9]

$$\hat{\mu}_{MAP} = \frac{n}{\tau+n} \cdot \bar{b} + \frac{\tau}{\tau+n} \cdot \nu \qquad (3)$$

$$\hat{\sigma}^2_{MAP} = \frac{\tau s^2 + nS^2 + \frac{\tau n(\bar{b}-\nu)^2}{\tau+n}}{\tau+n} \qquad (4)$$

where n is the total number of samples in the set $\{b_i\}$, $\bar{b}$ and S^2 are the sample mean and sample variance of that set, and τ, ν and s^2 are the hyper-parameters of $g(\mu,\theta)$.

2.2. Prior distribution estimation

Modeling the prior distributions at the level of phone units appears to be a good choice since phoneme is the basic unit of pronunciation. Although clustering allophones at the sub-phone level may characterize more details of accents, to reliably estimate the priors at such a level would require a significant amount of accent-specific training data. Assume that speech data of K speakers with accent Γ are given for estimating the priors. For each speaker k, the sample mean $e_{q,k}$ and variance $S^2_{q,k}$ of the data-model mismatch biases are first computed for each phone q. The hyper-parameters of the prior distribution for the phone q are then estimated as

$$v_q = \frac{1}{K}\sum_{k=1}^{K} e_{q,k} \text{ and } s_q^2 = \frac{1}{K}\sum_{k=1}^{K} S^2_{q,k} .$$

The hyper-parameters v_q and s_q^2 serve as the prior knowledge of data-model mismatch in phone unit q for the given accent.

2.3. Automatic accent detection and prior selection

The priors of bias distribution parameters are likely quite different for different foreign accents. This requires knowing the accent of each talker in order to apply proper priors in estimating the posterior bias distributions. On the other hand, due to mother tongue influence, speakers with the same foreign accent may consistently pronounce certain phonemes well and certain other phonemes poorly. This accent-specific pronunciation pattern of phonemes is reflected in the data-model mismatch over the defined phoneme set and can be utilized for automatic accent detection.

The accent models were trained in a similar way as the priors of bias distributions. For each accent Γ and each phone q, a Gaussian distribution $N(e_{\Gamma,q}, S^2_{\Gamma,q})$ is estimated from the bias samples in the training data, and the set of Gaussian distributions over the phone set becomes the accent model.

In testing, a set of biases $B = \{b_i\}$ are first calculated from the online test speech data. The average log-likelihood of B given Γ is then computed as:

$$\bar{L}(B \mid \Gamma) = \frac{1}{N}\sum_{q=1}^{Q}\sum_{j=1}^{N_q} \log\left[N(b_{q,j} \mid e_{\Gamma,q}, S^2_{\Gamma,q})\right] \qquad (5)$$

where Q is the number of phones, N is the total number of bias samples, N_q is the number of bias samples in phone q, $b_{q,j}$ is the

j-th bias sample in the phone q. The decision rule is

$$\Gamma^* = \arg\max_{\Gamma} \left[\overline{L}(B \mid \Gamma) + R(\Gamma) \right] \quad (6)$$

where $R(\Gamma) = C_{nat}$ if $\Gamma = \Gamma_{nat}$, and $R(\Gamma) = 0$ if $\Gamma \neq \Gamma_{nat}$, with Γ_{nat} denoting the native English speakers and $C_{nat} > 0$. The use of $R(\Gamma)$ in (6) is to reduce the risk of classifying a native English speaker to a foreign accent speaker (see section 3.2) .

The procedure of P-MEL based dynamic model selection and adaptation is similar to that of the MEL approach in [8], except that the bias distribution estimation is done under the MAP criterion, and accent-specific priors are determined by an accent detection algorithm described above.

3. EXPERIMENTAL RESULTS

3.1. Experimental condition

The baseline acoustic model was the same as used in [8] and was trained from the speaker-independent short-term training data (SI_TR_S, 200 speakers) of WSJ1. Within-word triphone HMM model each had three emitting states ("short-pause" model had a single state), and each state had a mixture of 16 Gaussian densities. Speech features consisted of 39 components of 12 MFCCs, energy, and their 1st and 2nd order derivatives. Cepstral Mean Normalization (CMN) as implemented in HTK was used. In testing, the 5K-vocabulary bigram language model of WSJ1 was used, and the decoder was provided by HTK v2.2 [10]. The acoustic model complexity and the decoding parameters of language model scale and word penalty were optimized for recognition of native speech [8].

In estimating the priors of bias distributions and performing recognition tests, data sets of native English speech, British accent speech and Chinese accent speech were used. Native English speech data consisted of WSJ1 set SI_DT_05 (NT1), and WSJ1 set SI_ET_H2 (NT2). NT1 and NT2 each had 10 speakers, with about 90 utterances per speaker in NT1 and 60 utterances per speaker in NT2. The British accent speech data came from LDC-WSJCAM0. Two groups were defined as BR1 and BR2, with each group having 10 speakers and about 85 utterances per speaker. The Chinese accent data set CH1 included three males and three females, where data collection were made under a similar acoustic condition and using the WSJ1 prompting texts. Each talker provided about 120 utterances. The baseline recognition performance on the five speaker groups is shown in Table 1. CH1 was the most difficult, with word error rate of 64.55%. British speaker groups also showed significant difficulties, with over 20% word error rate (absolute) than the native English speaker groups.

Table 1. Baseline recognition word error rates on the five data sets: NT1, NT2, BR1, BR2, CH1.

Group ID	NT1	NT2	BR1	BR2	CH1
Baseline WER %	10.86	9.67	31.41	35.62	64.55

In subsequent experiments, NT1 and BR1 were used to estimate the priors of the native speakers and British accent speakers, and NT2 and BR2 were used as the testing sets; the leave-one-out scheme was used for CH1, with the six speakers each used as the held-out test speaker once, and average word error was computed. In estimation of the hyper-parameters, the lower bound of the number of feature frames for estimating a bias sample was set to 35, and at least 30 samples of bias were used to estimate a prior distribution.

3.2. Analysis of prior distributions

Evaluations were made on model selection by using the estimated priors (without adaptation data) and its effect on recognition accuracy. Recognition results are summarized in Table 2 for the nine combinations of three test speaker groups and the three sets of priors. Compared with Table 1, Table 2 shows large performance improvements in BR2 and CH1, and only slight degradation in NT2. It can be observed from Table 2 that applying accent-mismatched priors to native English speakers caused significant performance degradation, whereas improvements were achieved for nonnative speakers even with mismatched priors. This result motivated the use of $R(\Gamma)$ in Eq.(6) to bias the detection output towards native speaker.

Table 2. Recognition word error rates after model selection by using the prior distributions only.

test set \ prior	CH1	BR1	NT1
CH1	54.05	54.19	57.75
BR2	32.34	32.07	32.66
NT2	11.15	11.51	10.40

3.2. Accent detection

Based on the priors of phones and phonetic classes, the accent detection method was evaluated on NT2, BR2, and CH1. Each speaker provided 40 adaptation utterances. For each test speaker, the first N adaptation utterances were used in accent detection, where N was set as 1, 3, 5, 10, 20, 40. When $N > 1$, the lower bound on the number of feature frames was set to be 35 for each bias sample, and when $N = 1$ the bound was set to 25. The constant C_{nat} was empirically set to be 2.5. Accent detection was performed according to (5) and (6).The detection error counts versus used utterances is shown in Table 3.

Table 3. Error count in accent detection

# utterances	1	3	5	10	20	40
CH1	1	1	0	0	0	0
BR2	0	0	0	0	0	0
NT2	1	0	0	0	0	0

3.3. P-MEL based dynamic model selection and adaptation

In implementing the P-MEL method, to avoid unreliable estimate of S^2 in (4), it was required that at least five samples of biases be accumulated in each phone or phonetic class, or else the bias distribution be estimated directly from the prior knowledge. For comparison, the MEL method as proposed in [8] was implemented under the similar condition, with the threshold on the number of biases for a full node set to be 25, and the threshold on the number of data frames for a full terminal GC set to be 30. Phonetic decision trees similar to those in [8] were used and the bias distributions were tied for prior sharing to have 42 clusters that corresponded to 42 phone units. In MAP bias

distribution estimation, τ was set to 15.

For each test speaker, the first N adaptation utterances, N = 1, 3, 5, 10, 20, 40, were used for model adaptation and selection, where the data partition for model selection and for model adaptation depended on N. When $N < 20$, all utterances were used in model selection, and a subset was used in initial model adaptation. Specifically, when $N = 1$, initial adaptation was not performed; when $N = 3$, the first utterance was used for initial adaptation to estimate a global bias-only transformation; when $N = 5$, the first two utterances were used for initial adaptation to estimate a global diagonal MLLR transformation; when $N = 10$, the first five utterances were used to estimate a global full MLLR transformation. For $N = 20$ or 40, half amount of adaptation utterances were used to estimate full MLLR transformations, where the sample size threshold for estimating a MLLR transformation was set to 500.

Recognition results on CH1, BR2, and NT2 are shown in Fig. 1. It is observed that, for native speech, there is no significant performance difference among the three approaches, but for nonnative speech, in every size of adaptation data, both MEL and P-MEL methods outperformed conventional MLLR method, and P-MEL performed better than MEL. The difference in error rates between MEL and P-MEL decreased with the number of adaptation sentences increased from 10 to 40. But the difference between the two methods was somewhat irregular with 1 to 5 adaptation sentences, possibly due to current choice of hyper-parameter τ and outliers in MEL that were associated with extremely limited data. It is worth noting that if the accent types of the foreign accent speakers were known, then the word error rate of the P-MEL method would be significantly reduced at $N = 0$, as given in Table 2 for the matched prior cases.

4. CONCLUSION

In this paper, a more reliable method for selecting model complexity in the case of sparse data is proposed by using maximum *a posteriori* estimation of bias distributions. Experiments were performed on American English speech, British accent speech, and mandarin Chinese accent speech. The results showed that the MAP approach enabled more reliable estimation of bias distributions with very small amounts of adaptation speech, or no adaptation speech. An automatic accent detection algorithm is also proposed for integration with dynamic model selection and adaptation. Recognition results show that accent detection combined with P-MEL is superior to the previous MEL method.

REFERENCES

[1] G. Zavaliakos, "Maximum *a posteriori* adaptation for large scale HMM recognizers," *Proc. ICASSP*, pp.725-728, 1996.

[2] U. Uelber and M. Boros, "Recognition of Non-native German Speech with Multilingual Recognizers," *Proc. of Eurospeech*, pp. 911-914, 1999.

[3] Witt, S. and Young, S., "Offline Acoustic Modeling of Nonnative Accents", *Proc. of Eurospeech*, 1999.

[4] S. Matsunaga, A. Ogawa, Y. Yamaguchi, and A. Imamura, "Non-Native English Speech Recognition Using Bilingual English Lexicon and Acoustic Models," *Proc. ICASSP'03*, vol I, pp 340-343, 2003

[5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[6] J. L. Gauvain and C. H. Lee, "Maximum *a Posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*., vol. 2, No.2, pp. 291-298, April 1994.

[7] X. He and Y. Zhao, "Model complexity optimization for nonnative English speakers," *Proc. of Eurospeech*, pp.1461-1464, Scandinavia, Denmark, September 2001.

[8] X. He and Y. Zhao, "Fast Model Selection Based Speaker Adaptation For Nonnative Speech," *IEEE Trans. on Speech and Audio Processing*. July 2003.

[9] M. H. DeGroot, *Optimal Statistical Decisions*, New York: McGraw-Hill, 1970.

[10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*, Version 2.2, http://htk.eng.cam.ac.uk/docs/docs.shtml.

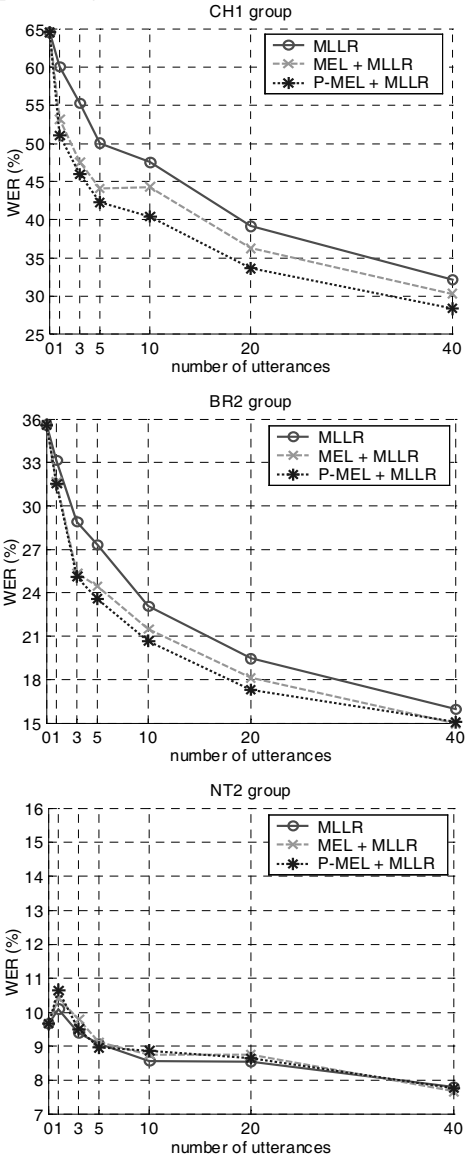

Fig. 1. Recognition WER vs. amount of adaptation data for Chinese, British, and native accent speaker sets CH1,BR2, NT2.