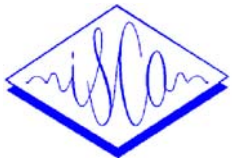

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

LINGUISTIC TREE BASED MAXIMUM LIKELIHOOD MODEL INTERPOLATION

Liu Feng, Chi-wei Che, Yu Peng, Wang Zuoying*

Speech Processing Lab., EE Depart., Tsinghua Univ., Beijing 100084, P.R.China

Email: *lf@thsp.ee.tsinghua.edu.cn*

*: Philips Innovation Center, Taipei, 24FA, 66, Sec. 1 Chung Hsiao W. Rd PO Box 22978, Taiwan

ABSTRACT

In this paper, a speaker adaptation method is presented which computes the speaker adapted model by a weighted sum of a set of speaker dependent models. The set of weights are estimated to maximize the likelihood of the adaptation data. Then a linguistic tree is constructed to cluster the mean vectors. The means in the same linguistic class share the same weight set, while the means in different classes use different weight set to compute the adapted model. Experiments show that with as little as 1~3 sentences a significant performance improvement is obtained. As more adaptation data is available, further improvement can be obtained.

10.21437/Eurospeech.1999-549

1 INTRODUCTION

In recent years, there has been much interest in speaker adaptation (SA) techniques for the large vocabulary continuous speech recognition (LVCSR) systems. It has been shown that the speaker adaptation is a very effective way to move the speaker independent(SI) performance to the speaker-dependent(SD) performance. Among the many SA schemes have been proposed, MAP[1] and MLLR[2] are the most promising. MAP has good asymptote performance, but it needs a large amount of adaptation data to estimate the model. MLLR uses a linear transformation to update the models, so it needs less adaptation data than MAP. However, MLLR also has a lot of parameters to estimate. For example, a global transformation has $d*(d+1)$ parameters (d is the dimension of the feature vector), then tens seconds to several minutes of speech data is required for the estimation. When the amount of adaptation data is too little (e.g., less than 10 seconds of speech), the performance may be even lower than that of a SI system[2][3].

As pointed in [4], speaker adaptation techniques should consider the following two issues: 1)speed of short term adaptation with only few data, and 2)asymptotic performance of long term adaptation with a large amount of adaptation data. As the asymptotic performance may be obtained by using MAP or the combination of MAP and other adaptation techniques [3,4], it becomes more important to investigate adaptation techniques which may move the SI models faster with little adaptation data. In addition, there are some cases where it is impossible to collect much speech for adaptation, for example, in the information retrieval system. In this case, it is more appropriate to use the adaptation schemes that may adapt quickly and efficiently.

Recently, some SA techniques are independently proposed which compute the SA model by making use of a set of SD models. [5] proposed a speaker clustering method which is based on finding a subset of speakers, from the training set, who are acoustically close to the test speaker, and then using only the training data from these speakers to re-estimate the SA models. [6] extended the idea of [5], which computes the SA model by an average of SD models that are closer to the test speaker. [7] proposed the reference speaker weighting (RSW) method, using a weighted sum of the set of SD models to compute the SA model. In [8], a similar technique named maximum likelihood model interpolation(MLMI) was proposed for the Mandarin large vocabulary continuous speech recognition system, and a significant performance improvement is reported with as little as 1~3 adaptation sentences. In [9], the author tried to construct a set of bases, called engenvoices, from the set of SD models, and represents the SA model by a linear combination of this set of engenvoices. In these methods, the parameters to be estimated are only a set of weights for all the SD models, so less adaptation data is needed. But on the other hand, the performance improvement saturates quickly as more adaptation data available, and this may be the main disadvantage of this sort of SA methods.

In this paper, we give an adaptation method named linguistic tree based maximum likelihood model interpolation (LT-MLMI), which has a significant improvement in performance given as little as 2~5 seconds of speech, and as more enrollment data is available, further improvement can be obtained. The basic idea of MLMI and the algorithm are described in section 2. In section 3, we discuss the MLMI with linguistic tree. Section 4 gives the experimental results and the conclusion is drawn in Section 5.

2 MAXIMUM LIKELIHOOD MODEL INTERPOLATION

The basic idea of MLMI is very simple[8]. Given a set of M speaker dependent models from the M corresponding training speakers, the speaker adapted mean vector is computed by the linear convex combination of the set of SD means and can be expressed as follows:

$$u_j^{(SA)} = \sum_{m=1}^{M} \alpha_m \cdot u_{mj}^{(SD)} \quad (1)$$

with the constraints:

$$\sum_{m=1}^{M} \alpha_m = 1, \qquad 0 \le \alpha_m \le 1, m = 1,...M \quad (2)$$

where $\{\alpha_m | m = 1,2,\ldots, M\}$ is the weights corresponding to the set of the SD models, and M is the number of training speakers in the training set;

$u_j^{(SA)}$ is the SA mean vector of state j. j=1,2,...,N, and N is the number of states in the SA models;

$u_{mj}^{(SD)}$ is the j-th mean vector of the SD model from the m-th training speaker, m=1,2,...,M,

Intuitively, if the characteristics of the test speaker is known, the SD models which have similar acoustic characteristics with the adaptation data should have larger weights, while others should have smaller weights. In the extreme case, if the test speaker is acoustically similar to one of the training speaker m, speaker dependent performance may be expected.

In MLMI, the parameters needed is the set of weights, and this number is relatively much less than MLLR and MAP (In our system, we use 152 SD models to compute the SA model), and hence faster adaptation may be expected.

3 MLMI WITH LINGUISTIC CLASSES

In previous section, we introduced the idea of MLMI, which tries to compute the SA model by making use of the inter-speaker relationship. While the MLMI has faster adaptation speed, its performance improvement saturates quickly when the amount of SA data increases. In this section, we investigate the methods for MLMI that can make full use of the adaptation data. In above case, a global combination is computed, i.e., all the means use the same weight set. As more adaptation data is available, we may use different weight sets for different mean vectors, so that the SA model may be approximated more accurately, and further improvement may be obtained.

The classification may be achieved using clustering of mixture components as [2]. In this paper, we cluster the mean vectors using a linguistic tree as [10]. Since the Chinese language is very different from English, we will first introduce the characteristics of Chinese language, and then construct a linguistic tree structure for MLMI. Finally, the estimation of the weights is given.

3.1 Characteristics of Chinese Language

The Chinese is different from the western language in three aspects:

1)The Chinese is not alphabetic, it is based on characters. The Chinese characters are monosyllabic. There are about 1254 syllables representing at least 100,000 characters.

2)Chinese is a tonal language, and every syllable or character is assigned a tone. There are 4 tones in our system. For example, a1, a2 denotes the syllable “a” with the first and second tone.

3)Conventionally, each syllable is decomposed into an “INITIAL/FINAL” format, in which “INITIAL” means the initial consonant of the syllable while the “FINAL” means the vowel or diphthong part, but including an optional medial or nasal ending. There are 22 INITIALs and 37 FINALs in Chinese.

For details of characteristics of Chinese language and speech, please refer [11].

3.2 The Linguistic Tree for MLMI

Based on the above observation, we use the FINAL-dependent INITIAL and the tone-dependent FINAL as the basic recognition unit in our system. To construct the linguistic tree, all the units are first divided into the INITIAL and FINAL classes. The INITIAL class contains all the FINALs and is further decomposed according to the broad phonetic subclasses[10]. Then each broad phonetic subclass is broken up into vowel-dependent consonants. The FINAL class is first decomposed into subclasses, each subclasses contains all the units that have the same vowels, for example, the “a” subclass contains the units of “ai”, “an”, “au”, etc. Then, each subclass is further decomposed into 37 FINAL sets, containing the units of the same FINAL with different tones. For example, the “a” subclass is decomposed into “ai”, “an”, “au”, etc., and the “ai” set contains the units “ai1”,”ai2”,”ai3” and “ai4”. In this way, we get a linguistic tree according to the characteristics of Chinese speech that shown in Fig.1.

3.3 Estimation of the Weight Parameters

In the adaptation process, the number of linguistic classes is determined according to the amount of the adaptation data, then all the SA mean vectors belonging to the same linguistic classes are computed using the same set of weights. The weights are estimated to maximize the likelihood of the speaker adapted models given the available adaptation data. For the r-th class, the weights are computed as follows:

$$a_r^* = C_r^{-1} b_r \qquad (r = 1,2,...,R) \qquad (3)$$

where R is the number of linguistic classes in MLMI, C_r is a M*M matrix belonging to the r-th linguistic class, with the component

$$c_{lk}^{(r)} = \sum_{i \in S_r} (u_{li}^{(SD)})' R_i^{-1} u_{ki}^{(SD)} \quad ,l,k=1,...,M \qquad (4)$$

b_r is a M-dimension vector, and

$$b_l^{(r)} = \sum_{i \in S_r} \frac{1}{T_i} \sum_{t=1}^{T_i} x_{it}' R_i^{-1} u_{li}^{(SD)} \quad \text{(l=1,2,..., M).} \quad (5)$$

In the above equations, x_{it} denotes the i-th frame of observations for state i, R_i is the i-th covariance matrix of the model, and S_r denotes the states that belong to the r-th class.

4 EXPERIMENTS

In the following experiments, we show the performance of the linguistic tree based MLMI, and we compare it with MLLR. To illustrate the effectiveness of the adaptation algorithm, only the acoustic performances are given. The baseline system has been described in[8], which is a scaled down version that used in the National Test for Mandarin Large Vocabulary Continuous Speech Recognition. The training data consists of speech of 152 speakers, 70 males and 82 females, each having from 500 to 600 sentences(about 40 minutes of speech). 152 SD models are used in MLMI. The gender dependent SI models are trained either using 70 males' speech or 82 females' speech. The test set consists of 10 speakers outside the training set. Each test speaker has 200 sentences for test and at most 400 sentences for adaptation. All the training and testing data are provided by the National 863 High-Tech Project.

In the first experiment, we show the performance of MLMI with few adaptation data. Tab.1 shows the syllable error rate of MLMI with respect to the number of adaptation sentences. All the means share the same set of weights. For comparison, we also give the performance of MLLR with a global full transformation matrix(MLLR-F) and MLLR with a global diagonal matrix(MLLR-D) in Tab.1. The baseline system has an average syllable error rate of 28.66% over the 10 test speakers. It was shown in Tab.1 that with as little as 1~3 sentences(2~5 seconds), the performance of MLLR-F becomes even worse than that of the SI system. The MLLR-D has a better performance than that with MLLR-F, which reduces the error rate to 25.82%, but the improvement saturates faster, and the asymptotic performance is very low. The MLMI shows an error rate of 23.26%, and outperforms the MLLR significantly. As the number of adaptation sentences increases, the performance of MLMI is worse than that of MLLR, this is because MLLR has much more parameters than MLMI.

#of sentences	MLMI	MLLR-D	MLLR-F
0(baseline)	28.66	28.66	28.66
1	25.96	26.80	87.85
2	23.53	25.96	51.11
3	23.26	25.82	36.74
5	22.82	25.63	28.36
7	22.73	25.46	25.17
10	22.95	25.71	23.97
20	22.84	25.37	22.00
40	22.70	25.44	21.56

Table1 MLMI with few adaptation data

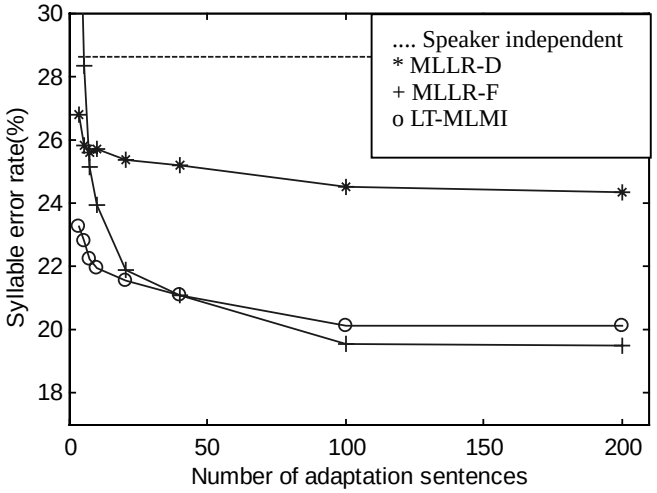

Fig.2 Performance of LT-MLMI

In the second experiment, we show the performance of linguistic tree based MLMI with different number of adaptation sentences. Supervised and batch adaptation is used. For MLMI, 152 SD model is used in adaptation, and the number of classes is chosen according to the amount of adaptation data. The MLLR uses the regression classes as in [2]. The syllable accuracy is shown in Fig.2. It may be seen from Fig.2 that the MLMI outperforms MLLR when the amount of adaptation data is little(less than 20 sentences in this experiment), while the performances of MLMI is very similar with that of MLLR for large amount of adaptation data.

5 CONCLUSIONS

How to utilize the speaker dependent model in the speaker adaptation is a very interesting topic. In this paper, we compute the SA model by the linear combination of the SD models. To be able to make full use of more adaptation data, we break the means of the model into different classes, and we use different sets of weights for different classes. In this way, further performance improvement is obtained.

ACKOWLEDGEMENT

The authors would like to express their appreciation to Dr. Roland Kuhn of Speech Technology Lab., Panasonic Technology Inc., for his many helpful suggestions and discussions on this work.

REFERENCES

[1] J.L.Gauvain and C.H.Lee, “Maximum a Posteriori estimation for multivariate Gaussian observations of Markov chains”, IEEE Trans. Speech and Audio Processing, vol.2, no.2, pp291-298, Apr.1994

[2] C.J.Leggetter and P.C.Woodland, “MLLR for speaker adaptation of continuous density hidden Markov models”, Computer Speech and Language, vol.9, no.2, pp171-186, 1995

[3] Vassilios V. Digalakis and Leonardo G.Neumeyer, “Speaker adaptation using combined transformation and Bayesian methods”, IEEE Trans on SAP, vol.4, no.4, July 1996, pp294-300

[4] Eric Thelen, Xavier Aubert, and Peter Beyerlein, “Speaker adaptation in the Philips system for large vocabulary continuous speech recognition”, Proceedings ICASSP’97, pp1035-1038, 1997

[5] M.Padmanabhan, L.R.Bahl, and M.A.Picheny, “Speaker Clustering and Transformation for Speaker Adaptation in Large-Vocabulary Speech Recognition Systems”, Proc. ICASSP-96

[6] Yuqing Gao, M.Padmanabhan, and M.Picheny, “Speaker Adaptation Based on Pre-Clustering Training Speakers”, Proc. ICASSP-98

[7] T.J.Hazen and J.R.Glass, “A comparison of novel techniques for instantaneous speaker adaptation”, Proceedings of Eurospeech’97, Sept. 97, Greece, pp2047-2050

[8] Wang Zuoying and Liu Feng, “Speaker adaptation using maximum likelihood model interpolation”, Proceedings ICASSP’99, also available at http://166.111.64.121.

[9] R.Kuhn, et.al, “Eigenvoices for speaker adaptation”, Proceedings of ICSLP’98, Dec. 1998, Austrilia, Paper number 303

[10] Prabhu Raghvan, and Chiwei Che, “Speaker Adaptation in Speech Recognition”, the Acoustical Society of America Meeting, December, 1997

[11] Lin-Shan Lee, “Voice dictation of Mandarin Chinese”, IEEE SP Magezine, July 1997, pp63-pp101

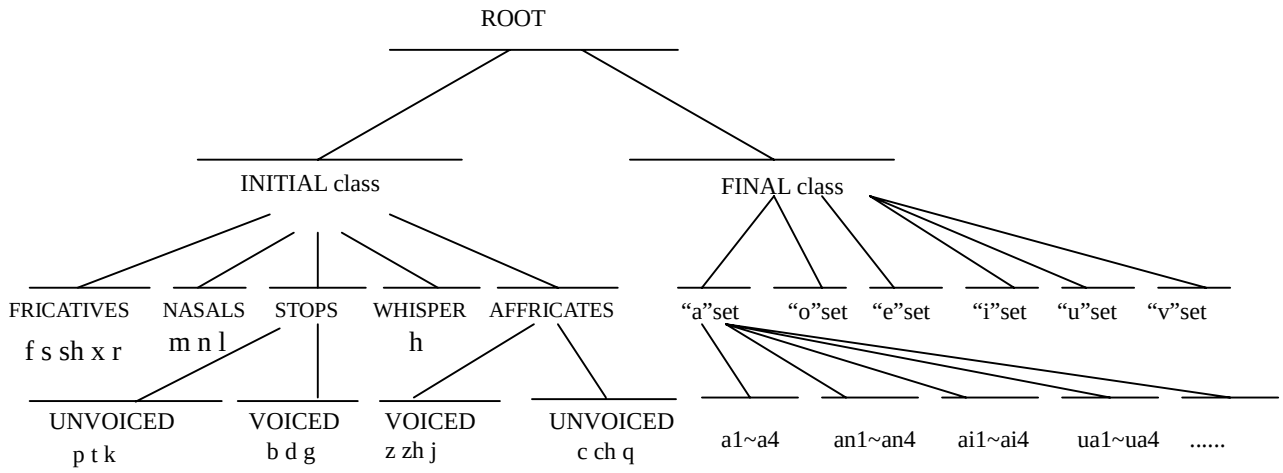

Fig.1 The structure of the linguistic tree

“a” set denotes the units that contain the vowel “a”, “an”, “ang”, etc.. In this figure, only a small fraction of the FINAL class is illustrated. "a1~a4” denotes the vowel “a” with different tones.