Optimization of Parameter-tying for Chinese Acoustic Modeling

Baosheng YUAN, Cuntai GUAN, Gareth LOUDON and Haizhou LI
Kent Ridge Digital Labs,
21 Heng Mui Keng Terrace, Singapore 119613
E-Mail: {baosheng, ctguan, gareth, hzli}@krdl.org.sg

ABSTRACT

Parameter-tying (or sharing) is widely used in hidden Markov models (HMM) for speech recognition because of its ability to enhance recognition accuracy and robustness. This paper tries to make best use of phonetic structure of the Chinese language to optimize parameter-tying in building acoustic models for an HMM-based continuous Chinese speech recognition system. Phonetic context based parameter-tying schemes are studied and compared with conventional parameter-tying schemes through experiments on a very large vocabulary, speaker-independent, continuous Chinese speech recognition system. The test of four male and four female subjects shows that the proposed parameter-tying scheme gives substantial improvement over the conventional ways of parameter-tying, without significantly increasing the number of system parameters.

I. INTRODUCTION

For a large vocabulary, speaker-independent, continuous speech recognition system, there is always a problem of insufficient training data. To overcome this problem, parameter-tying [Bahl 83] has emerged as a common practice since the early pioneering work based on discrete HMMs [Lee 88]. Parameter-tying has been more widely used recently in continuous density HMMs (CD-HMM) [Labiner 93] or semi-continuous HMM (SC-HMM) [Huang 90] for speech recognition and has been reported to improve recognition accuracy significantly [Vassilios 96], [Huang 90], Bellegarada 90].

There are mainly two ways to realize a parameter-tying scheme: data driven and phonetic knowledge based approach. While the data driven approach can be used in general to optimize parameter-tying automatically, the phonetic knowledge based method could be more precise in doing so for some languages (such as Mandarin Chinese) in which their phonetic/syllabic composition of a word is well constructed. This paper studies phonetic context based parameter-tying for Chinese acoustic models.

While the main purpose of this paper is to study parameter-tying, there is another related issue, i.e. selection of appropriate model units. This paper also addresses the latter issue through optimizing parameter-tying in building Chinese acoustic models.

In this paper, experiments on HMM-based speech recognition using conventional ways of parameter-tying are presented as benchmarks for comparison. Our work focuses on exploring an optimal way of parameter-tying for Chinese acoustic models based on phonetic knowledge of Chinese syllables. The parameter-tying in this work is at the HMM state level because it allows greater flexibility to achieve an optimal system performance. Since the phonetic transitional fragments of the syllable are more subtle and therefore more difficult to model than the “stationary” parts of the syllable, this paper focuses on how to model the transitional fragments of the syllable more accurately and optimize their parameter-tying scheme more effectively. In this study, we choose model units and construct the parameter-tying schemes on the models in such a way that the parameters representing the “stationary” fragments of the syllable are shared more efficiently; and the parameters representing the phonetic transitional fragments are tied more precisely.

In the next section, conventional ways of parameter-tying for HMM-based speech recognition are viewed and our approach is introduced. Section III gives detailed experiment results of the proposed parameter-tying schemes with comparison. Concluding remarks are given in the last section.

II. PARAMETER-TYING SCHEMES

2.1). Two unit models

2.1.1). Model units

Most research work uses the Initial and the Final of the syllable as basic models with different degree of context-dependencies[Hon 94], [Gao 95],[Lee 93], [Ma 96]. We use partial context dependent Initial and tonal Final to model a syllable. To deal with zero-Initial syllables, five additional pseudo-Initials are also used here as they have been shown to improve performance[Hon 94]. Each of the 27 Initials (21 lexical Initials and five pseudo-Initials) produces several medial (the beginning part of the Final)-dependent Initials, resulting in total number of 103 medial-dependent Initial models. Each tonal Final is treated as a unique model. This comes up with a total number of 250 Initial and Final models, each consisting of three states.

2.1.2). Conventional parameter-tyings

Between SC-HMM and CD-HMM, there are quite a few phonetics-based tying schemes. Phonetically tied HMM (PT-HMM) and phonetic state tied HMM (PST-HMM) are among the most popular names, therefore our work compares with these two schemes.

As one can see, PT-HMM (Figure 1,a) is not very precise in sharing the parameters among the models because it does not discriminate the different states of the models. The PST-HMM (Figure 1,b), on the other hand, although it treats the parameters of the different states separately, some state parameters are not properly tied. For example (in the Fig. 1, b), the third states of model /B_A/ and /B-E/ are tied together, which actually represent quite different phonetic segments; this therefore produces a less precise parameter-tying for those tied states.

2.1.3). Phonetic contextual state tied HMM

To amend the above inadequacies, we modified the PST-HMM by tying the third states of medial-dependent Initials separately to enhance the discriminative capability of the models. In this PCST-HMM scheme, we treat the different states of the Initial models differently. The first and the second states of the Initial models with same base phone (for example /B_A/, /B_E/ in the Fig. 1,c) are sequentially tied to their own mixture component sets respectively. The third states of the models with different context-dependencies are separately tied to their own mixture component sets. For the Final models, we simply use the PST-HMM scheme.

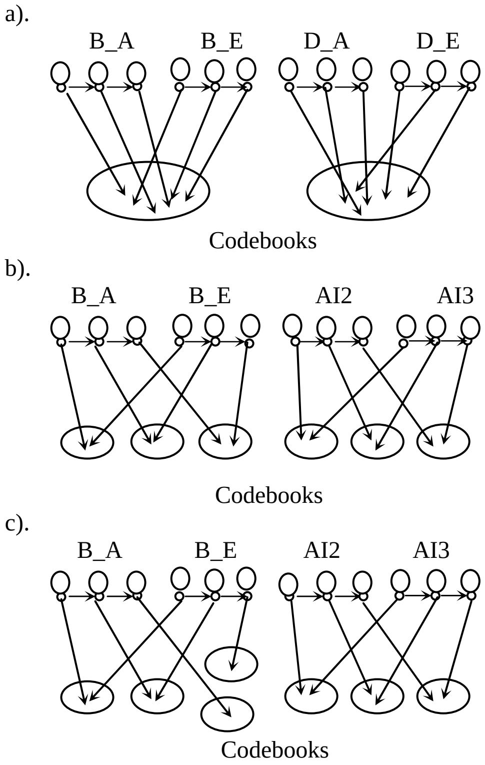

Figure 1. Parameter-tying schemes: a) PT-HMM; b) PST-HMM; c) PCST-HMM

2.2). Three unit models

2.2.1). Model units

For the three unit modeling, each syllable (/BAI3/, Figure 2) is represented by three sequential models: medial-dependent Initial /B_A/, transitional vowel /A(B,AI)/, and tonal Final /AI3/. By adding one more transitional model in the middle of a syllable, we are able to model the transitional fragments of the syllables more effectively and tying their parameters more accurately.

2.2.2). Parameter-tying (PCST-HMM-3)

The parameter-tying for all the states of Initial model and the Final model are the same as that in the two unit modling case, yet the parameter-tying scheme for each state of the transitional vowel model is dealt with separately (Fig. 2). The first state of the transitional model /A(B,AI)/ is tied with the last state of medial-dependent Initial model /B_A/; the second state of the model /A(B,AI)/ is tied with the first state of the tonal Final model /AI3/; and the last state of the model is tied to the second state of the tonal Final model.

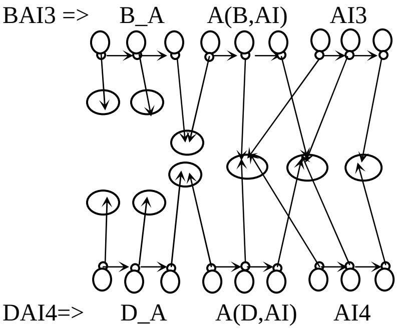

Figure 2. Three unit modeling PCST-HMM

III. EXPERIMENTAL RESULTS

3.1). Training data

Training data consists of 72 female speakers and 51 male speakers with each subjects recording about 80 minutes of speech, on the average. As a result, about 160 hours speech data was used to train gender-independent system.

3.2). Test data

The test subjects of four females and four males are used in the test, which are not in the training pool. The test set consists of 200 sentences for each of the eight subjects, which covers 90% of the model units. The character perplexity of the test sentence is about 100.

3.3). System parameters

The 13 element Mel-scale frequency Cepstrum coefficients (MFCC) and their first and second order differences were all used as Frontend features.

The following table compares the numbers of feature parameters for different parameter-tying schemes.

Table 3.1 Breakdown of Parameter Sizes

Tying scheme	Mods	Gauss	Cbooks	Mixture Wegihts
PT-HMM	250	64x65	65	3x250x64
PST-HMM	250	64x65x3	65x3	3x250x64
PCST-HMM	250	64x65x3 x3	65x3+76	3x250x64
PCST-HMM-3	250+ 440	64x65x3	65x3+76	3x(250+ 440)x64

3.4). Language models

In the speech recognition test, a lexicon of 50K words was used, which includes all 6763 GB characters and the most frequent 43238 words. The language model used in the test is a word class bigram based on 1000 word classes[Bai 98].

3.5). Test Results

The table below lists character/syllable recognition results (error rates) of eight speakers for the different parameter-tying schemes

Table 3.2. Chinese Character/Syllable (Error) Rate

	PT-HMM	PSTM-HMM	PCST-HMM	PCST-HMM-3
M1	24.8/23.0	20.4/18.7	18.7/17.1	18.5/17.1
M2	17.0/15.6	15.2/13.5	13.6/12.3	13.5/12.0
M3	21.2/19.4	20.0/18.5	17.7/16.1	17.2/15.2
M4	19.9/18.0	17.2/18.5	15.6/14.8	17.2/15.5
M-Av.	**20.7/19.0**	**18.2/17.3**	**16.6/15.1**	**16.6/15.0**
F1	16.9/14.5	14.0/11.8	13.1/11.2	13.3/11.4
F2	16.2/14.0	14.2/12.5	13.5/11.7	14.0/11.9
F3	19.4/17.8	15.9/14.0	14.5/12.9	14.9/13.3
F4	24.2/21.8	19.6/17.6	20.4/18.0	19.6/17.3
F-Av.	**19.2/17.0**	**15.9/14.0**	**15.4/13.5**	**15.5/13.5**
Ave	**20.0/18.0**	**17.1/15.6**	**16.0/14.3**	**16.0/14.2**

VI. CONCLUSION REMARKS

Phonetics-based HMM parameter-tying schemes for Chinese speech recognition have been studied and a phonetic contextual state parameter-tying scheme is introduced to optimize recognition performance. The proposed parameter-tying scheme paid more attention to more effectively model and tie the parameters representing the transitional fragments of a syllable to enhance their discriminating power to further improve the recognition rate. The experiments show that the proposed parameter-tying scheme substantially outperforms conventional phonetics knowledge based parameter-tying schemes without significantly increasing the number of system parameters. The proposed parameter-tying scheme can be further extended to more sophisticated context-dependent Chinese acoustic models such as cross-syllable context-dependent acoustic models so long as the training data is sufficient for such detailed models.

REFERENCES:

[Bai 98] Bai Shuanhu, Li Haizhou, Li Zhiwei and Yuan Baosheng, "Building class-based language models with contextual statistics", Proc. ICASSP'98, pp.173-76, Seattle Whashington, USA, 1998

[Bellegarada 90] J. R. Bellegarda, and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition", IEEE Trans. Acoust., Speech, Signal Processing, Vol.38, pp.2033-2045, Dec., 1990

[Bahl83] L.R.Bahl, F. Jelinek, and L. R. Mercer, "Amaximum likelihood approach to continuous speech recognition", IEEE Trans. Pattern Anal. Machine Intell., PAMI-5, pp.179-90, 1983.

[Gao 95]Yuqing Gao, Hsiao-Wuen Hon, Zhiwei Lin, Gareth Loudon, S. Yoganathan and Baosheng Yuan, "Tangerine: A large vocabulary mandarin dictation system", Proc. IEEE ICASSP'95, pp77-80, Detroit USA, May 1995.

[Hon 94] Hsiao-Wuen Hon, Bao-Sheng Yuan, Yen-Lu Chow, Shankar Narayan and Kai-Fu Lee, "Towards large vocabulary mandarin Chinese speech recognition", Proc. IEEE ICASSP'94, pp. 545-548, Adelaide, Australia, April 1994.

[Hon 97] Hon, Hsiao-Wuen and Yuan, Bao-Sheng, "System and method for generating and using context dependent sub-syllable models to recognize a tonal language", U.S Patent, No. 5680510, Oct. 21, 1997.

[Huang 90] X. D. Huang and M. A. Jack, "Semi-continuous hidien Markov models for speech recognition", in Reading in Speech Recognition, A. Waibel and K.F. Lee, Ed. New York: Morgan Kaufmann, pp.340-346, 1990

[Ma 96] Bin Ma, Taiyi Huang, Bo Xu, Xijun Zhang and Fei Qu, "Context-dependent acoustic models for Chinese speech recognition", Proc. IEEE ICASSP'96, pp.455-558, Atlanta,USA, 1996

[Labiner93] Lawrence Labinerand and Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice Hall PTR, Englewood Cliffs, NewJersey, 1993.

[Lee 88] Kai-Fu Lee, "Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system", Ph.D Dissertation, Computer Science Dept., Carnegie Mellon University, USA, April 1988.

[Lee 93] Lin-Shan Lee et al, "Golden Mandarin (I) - A real-time mandarin speech dictation machine for Chinese language with very large vocabulary", IEEE Trans. Speech and Audio Processing., Vol. 1, No.2, pp.158-179, April 1993

[Vassilios 96] Vasilios. V. Digalakis, Peter Monaco, and Hy Murveit, "Genones: generalized mixture tying in continuous hidden Markov model-based speech recognition", IEEE Trans.Speech & Audio processing, Vol.4, No.4, July 1996.

[Yuan 97] Yuan Baosheng, "A very large-scale continuous speech database for mandarin Chinese" ISS Internal Technical Report, National University of Singapore, May 1997.