DATA-DRIVEN ACOUSTIC MODELING APPROACH FOR CHINESE LVCSR

Cuntai GUAN, Haizhou LI, Baosheng YUAN, Zhiwei LIN
Kent Ridge Digital Labs,
21 Heng Mui Keng Terrace, Singapore 119613
E-mail: {ctguan, hzli, baosheng,lzw}@krdl.org.sg

ABSTRACT

In this paper, we propose a data-driven approach for building the acoustic models of a LVCSR system. It can automatically select the best acoustic models in a maximum likelihood framework. This algorithm jointly finds out a model mapping function and estimates the model parameters with expectation/conditioned maximisation (ECM), which is an extension of the conventional EM algorithm. This approach is evaluated on a Chinese LVCSR system and similar performance is obtained as compared to its phonetics-driven counterpart. This method can be easily adopted by a multi-lingual speech recognition system.

1. INTRODUCTION

For a large vocabulary, continuous speech recognition (LVCSR) system, the acoustic models are among the most important components [1]. Many systems adopt acoustic models defined and finely tuned based on phonetics and linguistics clues of the specific languages, as called phonetics-driven models in this paper. It is usually not trivial to re-scale the number of models according to new task. Recently, multilingual speech recognition raised great interest among researchers [3,4]. One of the important properties of a multilingual speech recognition system is that, it can be easily adapted to a variety of applications for different languages, hopefully with less effort. The multi-lingual speech recognition is of particular interest in Singapore due to its multilingual environment. In a LVCSR system, the commonly used types of models are context-dependent phones, such as triphones. In Mandarin speech recognition system, initial and final models are also among the most popular acoustic models [5]. Aiming at building up a general method for multilingual speech recognition system, we chose to study the acoustic models defined with triphones within a word (or syllables for Mandarin).

In this paper, we focus on the acoustic model issue of the multilingual speech recognition. The motivations for the work presented in this paper are two fold. First, in a real-world application, an automatic procedure to select a smaller number of context-dependent models from a large pool is often required, especially for a real-time recognition system, where the computation complexity is a great concern due to the current computing constraint. Second, this method can be used in the multilingual LVCSR system.

We propose here a data-driven approach to cope with these two problems. The method presented in this paper is composed of a model pre-select procedure and a mapping/estimation phase in a maximum likelihood framework. First, a subset of acoustic models is selected from a large corpus of models based on maximum likelihood criterion. Then, the derivation of a mapping function, where the mapping function one-to-one maps each original model to a destination model, and the re-estimation of the destination models are carried out in a estimation/conditioned maximisation (ECM) framework. ECM is an extension of the conventional EM algorithm [6] and has been successfully applied in [2] to jointly estimate the values of the model parameters and the tying structure of the HMM models. The work in this paper can be considered as the further extension of [2] from state mapping to model mapping. The concept of model mapping is illustrated in Fig 1.

This approach was evaluated on a Chinese LVCSR system. Its performance was compared with our phonetics-driven Chinese LVCSR system, which is considered as a benchmark. The experimental results showed that the data-driven method

produced similar results to the phonetics-driven one with only a slight degradation in performance.

In the next section the algorithm of the proposed approach is introduced, followed by section 3 in which some practical issues are addressed. Experimental results are presented in section 4. Conclusion will be given in the last section.

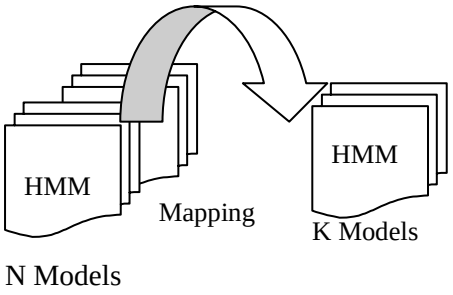

Figure 1 Mapping from *N* models to *K* models

2. DATA-DRIVEN ACOUSTICAL MODELING APPROACH

Given an acoustic model, $\omega_n \in (\omega_1,...,\omega_N)$, represented by a hidden Markov model, and the training data consisting of an observation sequence $O^{(n)} = (o_1^{(n)},...,o_T^{(n)})$ of T feature vectors generated by the model ω_n, the conventional EM algorithm for training the models includes, a maximization step to maximize probability $P(O^{(n)} / \omega_n)$, and an estimation step to estimate the model parameters for each of the *N* models. Here the purpose of this work is to select a set of *K* models $\Xi = \{\xi_1,...,\xi_K\}$ from *N* initial acoustic models $\Omega = \{\omega_1,...,\omega_N\}$ with the EM algorithm, where $K \le N$ and $\Xi \subset \Omega$. There exists a mapping function between the initial acoustical models Ω and the destination models Ξ:

$$\eta(\,) : \{1,.n., N\} \to \{1,.k., K\} \tag{1}$$

To find out this mapping function, we choose to maximize the following probability, instead of maximizing the individual probability $P(O^{(n)} / \omega_n)$ for each model as in the conventional EM training procedure:

$$\sum_{k=1}^{K} P(O^{(k)} / \xi_k) = \sum_{\{n|\eta(n)=k\}} P(O^{(n)} / \omega_n) \tag{2}$$

The maximization of (2) can be carried out straightforwardly in the similar way as in the Balm-Welch algorithm[1], as there are only a finite number of distinct mapping combinations, say $N \times K$. However, the computation cost will become quite large. To reduce the computational complexity, we adopt the estimation/conditioned maximization (ECM) algorithm [2,6]. The principle of ECM consists of replacing the maximization step of EM with a sequence of constrained maximization with respect to different steps of parameters. The convergence property of ECM was analyzed in [2] and [6]. The training procedure is thus divided into two successive processes, a models pre-select process and an estimation/conditioned maximization process.

A. Model Pre-select Process

During this phase, the values of all the model parameters are held unchanged. *K* models are selected according to the criterion in (3). Supposing Γ a subset of the initial model indices, $\Gamma \subset \{1,..., N\}$ and rank(Γ)=*K*, from the *N* models, one can select *K* models whose indices are γ_k, $\gamma_k \in \Gamma$:

$$\Gamma = \{\gamma_1,..\gamma_k.,\gamma_K\} = \arg \max_{\substack{\Gamma' \\ rank(\Gamma')=K}} \sum_{i\in\Gamma'} P(O^{(i)} / \omega_i) \tag{3}$$

B. ECM Process

This phase is composed of two iterative processes. First, a mapping function is updated through the following maximization process by assuming the Gaussians fixed,

$$\eta(n) = \arg\max_k P(O^{(n)} / \omega_{\gamma_k})$$
for $n=1,\ldots,N$ (4)

Then, the HMM parameters of each model are estimated. The estimation method is generally the same as the commonly used methods[1] together with the above derived mapping function. If the gradient method is deployed to estimate the parameters, the gradient function now becomes:

$$\frac{\partial}{\partial \lambda} P(O^{(k)} / \omega_k) \quad (5)$$

where λ represents the model parameter. $O^{(k)} = \bigcup_{\{n|\eta(n)=k\}} O^{(n)}$.

3. SOME PRACTICAL ISSUES IN APPLICATIONS

There will be some practical issues in realizing the above algorithm. First, to use the method introduced in 2 in the explicit way, the computational complexity would be proportional to $N \times K$. In a Chinese LVCSR system, the number of the initial models are over 1000, therefore, it is very computationally consuming to accomplish the training procedure in practice. Second, in the above procedures, since the mapping function is estimated with the "initial" models at the beginning phase of the training procedure only, it may not be precise enough to reflect the acoustical similarity of the models.

To solve the first problem, we deploy the following scheme to reduce the computational complexity. All the initial models are roughly divided into subsets, which merely makes use of certain rough phonetics knowledge. As such, for each subset, the number of initial models *N* is reduced to several or tens.

To tackle the second problem and make the mapping function more reliable, we carry out the training process step-by-step. During the pre-select step, the number of destination models is set to be equal to the number of initial models minus one. After the ECM process finishes, then a new set of destination models is selected by setting the resultant models in the previous procedure as the new initial models. These same processes are repeated until reaching a preset destination model number.

The experimental results for a Chinese LVCSR will be discussed in the next section.

4. EXPERIMENTAL RESULTS

The proposed approach was evaluated in a speaker-independent Chinese LVCSR system for dictation style. The training database contains 51 male speakers, with totally 68 hours of continuous sentences. The test data is from 4 male speakers, each of whom contributes 225 sentences on average. The vocabulary of the system consists of 50K words. N-gram statistics are used for language modelling. The acoustic parameters consist of 13 mel cepstral coefficients with delta and delta-delta parameters. The recognizer makes use of continuous density HMM with Gaussian mixture distributions. Each model is a 3 state left-to-right HMM model.

Three different kinds of acoustic models are tested. The first is the phonetics-driven one consisting in 315 distinct models with state tie (denoted as PDM), which is considered as the benchmark for comparison. The second set of models is composed of 1690 triphones (denoted as TPM). The last set of models consists of 323 models which are derived from the 1690 triphones with the aforementioned data-driven method (denoted as DDM). All the three sets of tests use the same training data, the same vocabulary, and the same language model. The independent model parameters for each of these three sets of acoustic models and the tonal syllable recognition accuracy are listed in Tab 1. The result shows that when the number of independent parameters drops from 12.8M of the TPM to 1.25M, the DDM has the similar recognition accuracy to PDM with only a slight degradation in performance. In our experiment, some scheme may need to be further adjusted. For instance, the choosing of number of destination models for each subset adopts a simple logarithm function, which may be not adequate for some phones, such as consonants. The fine tune of such a scheme may further improve the performance.

Table 1. Recognition Accuracy

MODEL TYPE	ACC. %	NO. PARA (M)
PDM	83.5	1.03
TPM	84.2	12.8
DDM	83.2	1.25

5. CONCLUSION

The approach proposed in this paper provides a feasible way of automatically scaling down the number of context-dependent models in a HMM based LVCSR system. No delicate linguistics or phonetics knowledge is needed in this method and the resulted models are reliable, as far as the training data is somewhat sufficient as is often the

case since many large speech data corpora have been developed. Although this approach was only evaluated on a Chinese LVCSR system, we believe it is independent of language and therefore can be easily adopted to other languages so as to build a multilingual speech recognition system.

References

1. Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.
2. Olivier Cappe, Chafik E. Mokbel, Denis Jouvet and Eric Moulines, "An algorithm for maximum likelihood estimation of hidden Markov models with unknown state-tying", *IEEE Trans. Speech and Audio Processing*, Vol.6, No.1, Jan, 1998.
3. Li Deng, "Integrated-Multilingual Speech Recognition Using Universal Phonological Features In A Functional Speech Production Model", IEEE Proc. ICASSP'97, pp1007-1010.
4. L. Lambel, M.Adda-Decker, J.L.Gauvain, "Issues in Large Vocabulary, Multilingual Speech Recognition", Proc. EUROSPEECH-95, pp1-4.
5. Lin-shan Lee, "Voice Dictation of Mandarin Chinese", IEEE Singal Processing Magazine, July, 1997, pp63-101.
6. X.L.Meng, D.B.Rubin, "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework", Biometrika, Vol.80, pp267-278, 1993.