SPEAKER-ENSEMBLE HIDDEN MARKOV MODELING FOR AUTOMATIC SPEECH RECOGNITION

Guoli Ye and Brian Mak

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
{yeguoli, mak}@cse.ust.hk

This paper proposes a new hidden Makov model (HMM) which we call *speaker-ensemble HMM* (SE-HMM). An SE-HMM is a multi-path HMM in which each path is an HMM constructed from the training data of a different speaker. SE-HMM may be considered a form of template-based acoustic model where speaker-specific acoustic templates are compressed statistically into speaker-specific HMMs. However, one has the flexibility of building SE-HMM at various level of compression: SE-HMM may be built for a triphone state, a triphone, a whole utterance, or other convenient phonetic units. As a result, SE-HMM contains more details than conventional HMM, but is much smaller than common template-based acoustic models. Furthermore, the construction of SE-HMM is simple, and since it is still an HMM, its construction and computation is well supported by common HMM toolkits such as HTK. The proposed SE-HMM was evaluated on Resource Management and Wall Street Journal tasks, and it consistently gives better word recognition results than conventional HMM.

Index Terms— detailed acoustic modeling, template-based automatic speech recognition, speaker-ensemble acoustic model

1. INTRODUCTION

The success in automatic speech recognition (ASR) is partly attributed to the development of context-dependent acoustic modeling. One common technique in context-dependent acoustic modeling is *parameter tying*, which enables acoustic models to be trained robustly for a given (perhaps limited) amount of training data, and yet the models are compact and decode at reasonable speed. Different model parameters have been tied, resulting in generalized triphone models [1], semi-continuous hidden Markov model (HMM) [2], tied-state HMM [3], subspace distribution clustering HMM [4], and so forth.

In recent years, with the availability of large amount of training data and the advance of computer technology (e.g., large but cheap memory, faster CPU), there are researches in the opposite direction to build more detailed acoustic model. In theory, more detailed acoustic model could model the acoustic-phonetic characteristics more accurately. One representative example is the template-based acoustic model (TBAM) [5, 6], which shows very competitive results when compared with traditional HMMs. To build a template-based triphone acoustic model, a set of speech templates of the triphone are first collected from the training data. In the most common case, each template is simply a sequence of feature vectors that represents an actual occurrence of the triphone in the corpus. As the number of templates for a triphone could be quite large, clustering or selection technique is usually used to reduce the large number of triphone templates into a manageable number of template representatives, which are supposed to store the fine phonetic details of the triphone. This is different from an HMM-based ASR system, where each triphone is usually represented by a highly abstract HMM with a set of (parametric or non-parametric) probability distributions.

Despite the ability to characterize the phonetic details, template-based acoustic modeling has its own problems. Firstly, its model size is much larger than HMM as it needs to store a lot of template representatives. Secondly, many of the useful techniques developed for HMM, e.g., speaker adaptation, discriminative training, may not be easily applied to TBAM. Thirdly, for beginners, it is not easy to build TBAMs by themselves as there are not many well-established tools to support it, while for HMM-based acoustic modeling, many well developed toolkits such as HTK [7] and Kaldi [8] are freely available.

In this paper, we propose a new acoustic modeling method called *speaker-ensemble HMM* (SE-HMM) with the aim to combine the advantages of traditional hidden Markov modeling and template-based acoustic modeling. When our new method is applied at the phone level to create SE triphone HMMs, each triphone is represented by a mixture of traditional 3-state strictly left-to-right HMMs, with each HMM "component" describing a particular speaker's realizations of the triphone. The number of HMM components in an SE tri-

phone HMM is equal to the number of speakers in the training corpus. From the perspective of template-based acoustic modeling, each HMM component could be viewed as a statistical template representative of the triphone from a specific speaker. Compared with traditional HMM, SE-HMM stores more acoustic-phonetic details as in common TBAMs. On the other hand, SE-HMM has the following advantages over common TBAM:

- The association of each training speaker with an HMM may be seen as a natural way to cluster acoustic templates from each speaker in the TBAM approach.
- The model size is smaller. As will be seen in the next section, only the speaker-independent (SI) HMM parameters and speaker-dependent (SD) transformation matrices need to be stored.
- Since a mixture of HMM components is still an HMM, the final SE-HMM is an HMM. Thus, all HMM operations such as speaker adaptation and discriminative training techniques that work for conventional HMM may be applied to SE-HMM.
- Common HMM tools such as HTK and Kaldi can be readily employed for the construction of SE-HMM.

Multi-path HMM has been tried in the past. In [9], senone-dependent speaker-clustered HMMs are built by clustering speaker data at the senone (tied state) level into at most 8 clusters per senone. In [10], syllable-length acoustic trajectories are clustered to construct multi-path HMMs for the 94 most frequent syllables, which are then mixed with traditional triphone HMMs in a large-vocabulary continuous speech recognition (LVCSR) task in Dutch. Similarly, Korkmazskiy proposed the generalized mixture of HMMs [11] for a continuous digit recognition task. Unlike all these works which require some kind of clustering of speech sequences, we simply keep all speaker acoustic characteristics in our SE-HMM in a succinct manner using their HMMs in the spirit of template-based acoustic modeling. Furthermore, our modeling approach is very flexible and can be applied to construct SE-HMM states, SE triphone HMMs, SE syllable HMMs, and so forth. As will be seen in Section 3, our experiments in LVCSR show that SE triphone HMMs perform better than SE-HMM states or SE utterance HMMs.

2. SPEAKER-ENSEMBLE HMM (SE-HMM)

Speaker-ensemble hidden Markov modeling is very flexible. In this paper, we will investigate its construction at the subphonetic level (state), phone level, and utterance level. Without loss of generality, we will describe the construction of speaker-ensemble triphone HMM below; speaker-ensemble HMM state and speaker-ensemble utterance HMM can be constructed in a similar procedure.

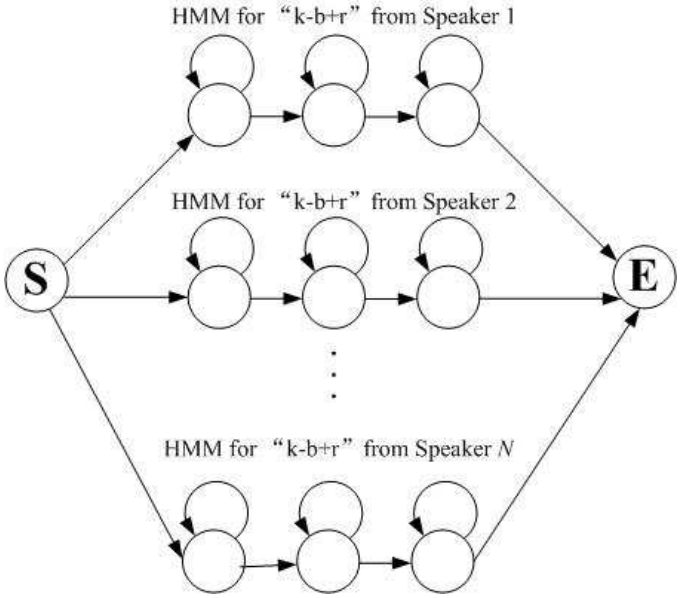

Fig. 1. A multi-path HMM representing the SE-HMM for the triphone "k-b+r", where N is the number of training speakers, and nodes **S** and **E** are the non-emitting entry and exit nodes.

2.1. Construction of Speaker-ensemble Triphone HMM

The procedure to construct SE triphone HMM is described as follows.

STEP 1: Train a baseline speaker-independent (SI) triphone HMM system.

STEP 2: Build speaker-dependent (SD) models for each speaker in the training data by adapting the SI models with MLLR [12].

STEP 3: For each triphone p, collect all its SD-HMMs, $\lambda_{p,k}^{(SD)}$ from each of the training speakers $k = 1, \ldots, N$.

STEP 4: To construct the SE-HMM $\lambda_p^{(SE)}$ for triphone p, a multi-path triphone HMM is then composed from the set of SD triphone HMMs, $\{\lambda_{p,k}^{(SD)}, k = 1, \ldots, N\}$ with one SD triphone HMM per path. The transition probabilities from the non-emitting entry node to the multiple paths are set proportionally to the amount of training data for the training speakers. Figure 1 gives an example of a multi-path SE-HMM that may represent the triphone "k-b+r".

In practice, the number of training speakers N is usually not small. For example, there are 109 and 83 speakers in the Resource Management and Wall Street Journal corpora respectively. If we store all the SD-HMM parameters as shown in Figure 1, the size of the resulting SE-HMM could be very large. To save space, one may store only the SI models and all the SD MLLR transformation matrices. During decoding, the SD-HMM parameters can be generated on-the-fly using their

SI counterparts together with the appropriate MLLR transformation matrices.

2.2. Augmented SE-HMM

We also investigate a variation of the SE-HMM which we call *augmented SE-HMM*. An augmented SE-HMM has $N + 1$ paths: besides the N SD-HMMs mentioned above, an additional path that represents its SI model is added. The transition probability from the entry node to the SI-HMM path is set to 0.5, while the transition probabilities of the other N SD-HMM paths are rescaled so that they sum to 0.5.

The idea of introducing the additional SI path in the augmented SE-HMM is that in case the phonetic realization of a triphone in the testing data matches poorly with any of the SD-HMM paths, it may match better with the SI path. In other words, when the acoustic data from a testing speaker seriously mismatch with any of the training speakers, it is better to back-off to the baseline SI model.

2.3. Decoding of SE-HMM

Since an SE-HMM is just another HMM, it can be used directly in speech decoding by common recognition tools such as HTK. However, as the models are much more complex — N times larger — than the SI models, it will be very time consuming to decode with them directly. In practice, 2-pass decoding is adopted: in the first pass, a lattice is generated by the SI models; in the second pass, the SE-HMMs are used to rescore the lattice to get the final recognition outputs.

3. EXPERIMENTAL EVALUATION

The proposed SE-HMM was evaluated on the Resource Management (RM) and Wall Street Journal (WSJ) tasks. In both tasks, the conventional 39-dimensional MFCC vectors (with energy, delta, and delta-delta) were extracted at every 10ms over a window of 25ms. Cross-word triphone models were then constructed using the HTK toolkit. The baseline speaker-independent (SI) models are strictly left-to-right 3-state continuous-density HMMs (CDHMM) with a Gaussian mixture density at each state. In addition, there are a 1-state short pause model and a 3-state silence model.

All system parameters such as the decoding parameters, state-tying tree, and the number of regression classes for MLLR adaptation were optimized using their respective development data set. All recognition results are reported in word error rate (WER).

3.1. Resource Management Task

3.1.1. Speech Corpus

The RM system was built from the standard SI-109 training data which consist of 3,990 utterances from 109 speakers, and was evaluated on the four standard test sets: Feb'89, Oct'89, Feb'91 and Sep'92 test sets using the standard word-pair grammar. The speaker-dependent development set consisting of 1,200 utterances was used to tune the various system parameters such as the state-tying tree and the decoding parameters.

3.1.2. Experimental Setup

The baseline SI system consists of 6,817 cross-word triphone HMMs. Each triphone state has a Gaussian mixture density of at most 6 components. There are totally 1,589 tied states which were derived from a phonetic decision tree.

109 sets of SD triphone models were adapted from the set of SI models by MLLR, one per training speaker. A binary regression tree of 16 classes was used in the MLLR adaptation. SE triphone HMMs and augmented SE triphone HMMs were constructed from the corresponding SD triphone HMMs and SI triphone HMM by following the procedure described in Section 2.

3.2. Wall Street Journal Task

3.2.1. Speech Corpus

The standard SI-84 Wall Street Journal (WSJ0) training set was used for training the speaker-independent models. It consists of 7,138 utterances from 83 speakers for a total of about 15 hours of read speech. All the training data were end-pointed. The standard Nov'92 5K non-verbalized test set was used for evaluation using the standard 5K-vocabulary bigram that comes along with the WSJ corpus. The development set si_dt_05, containing 409 sentences, was used to tune the system parameters.

3.2.2. Experimental Setup

The baseline SI system consists of 12,581 cross-word triphones. Each triphone state has a Gaussian mixture density of at most 16 components. There are totally 2,796 tied states which were derived from a phonetic decision tree.

Eight-three sets of SD models were adapted from the SI models again by MLLR, one per training speaker. Sixteen regression classes were used in the MLLR adaptation. The SD models were then used to build the SE-HMMs and augmented SE-HMMs as in the RM task.

3.3. Performance of Various Acoustic Models

The performance of different models on both RM and WSJ0 tasks is shown in Table 1. For each model, we tried our best effort to figure out its optimal setting. For the RM task, the reported WER is an average over the four test sets.

From Table 1, it can be seen that SE-HMM performs consistently better than the baseline SI-HMM with a relative

Table 1. Recognition performance (WER) of various models.

Model	RM	WSJ
Baseline SI-HMM	3.83%	6.54%
SE-HMM	3.60%	6.20%
Augmented SE-HMM	3.57%	6.28%

Table 2. Recognition performance (WER) of SE-HMM at different level of speaker consistency.

Consistency Level	RM	WSJ
state	3.80%	6.59%
phone	3.60%	6.20%
utterance	3.89%	6.84%

WER reduction of 6% and 5% on the RM and WSJ tasks respectively. Moreover, the augmented SE-HMM does not perform better than SE-HMM. We suspected that the SI-HMM path in the augmented SE-HMMs was seldom taken in the Viterbi search. To verify that, we analyzed the recognized state sequences from the augmented SE-HMMs on the WSJ task and found that the SI paths were only chosen for 4% of the time. The small usage of the SI paths probably would not affect the system performance that much.

3.4. Different Levels of Speaker Consistency

From the template-based acoustic modeling's point of view, SE triphone HMM stores speaker-specific template representatives (SD-HMMs) at the phone (or triphone) level. As a result, the decoded path for a testing utterance consists of a sequence of triphones with the following "phone-level speaker consistency" property: on a decoded path, different triphones may come from different speakers but the states within a triphone must come from the same speaker. In the following, we would like to investigate the importance of this property by building SE-HMMs at the state level and the utterance level.

3.4.1. Utterance-level Speaker Consistency

In utterance-level speaker consistency, any speech segment in the whole decoded path of a testing utterance must come from the same speaker. This requirement can be easily achieved by decoding each testing utterance N times, each time with one of the N MLLR-adapted SD models. The SD model which gives the highest recognition score is chosen as the final model to decode the utterance.

3.4.2. State-level Speaker Consistency

In state-level speaker consistency, each state in the decoded path is allowed to come from a different speaker. The construction of SE-HMM triphone states to fulfill this property is similar as the construction of SE triphone HMMs, except that the parallel paths are combined at the state level instead of phone level.

3.4.3. Comparison Results

Table 2 compares the performance of requiring speaker consistency at various modeling levels. It is observed that the requirement of speaker consistency at both the state and utterance levels results in poorer performance than that at the phone level. In fact, their performances are not even better than the baseline SI model.

The results seem to suggest that the requirement of speaker consistency over a whole utterance is too restrictive. Given the limited number of training speakers, it is unlikely that a test speaker will match one of the training speakers perfectly over all triphones. On the other hand, the requirement of speaker consistency at the state level seems too flexible, and suffers from the trajectory folding problem [10] of allowing switching between speakers among the three triphone states. The requirement of speaker consistency over a triphone thus strikes a good balance between flexibility and consistency.

4. CONCLUSION AND FUTURE WORK

In this paper, a novel acoustic model called speaker-ensemble HMM (SE-HMM) is proposed to combine the advantages of traditional HMM and template-based acoustic model. SE-HMM contains more speaker acoustic details than traditional HMM. It can also be considered as a compressed version of the template-based acoustic model in which the acoustic templates from each training speaker are compressed into an HMM. Since SE-HMM is an HMM, it can be readily implemented and manipulated using current HMM tools.

Recognition experiments on the RM and WSJ0 tasks show that SE-HMMs perform better than traditional HMMs on both tasks if it is implemented to maintain speaker consistency at the phone level (instead of state or utterance level).

Future works include comparing SE-HMM with other template-based acoustic models and speaker-clustered models.

5. REFERENCES

[1] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 4, pp. 599–609, 1990.

[2] X. D. Huang and M. A. Jack, "Semi-continuous hidden Markov models for speech signals," *Computer Speech and Language*, vol. 3, no. 3, pp. 239–251, 1989.

[3] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 307–312.

[4] E. Bocchieri and B. Mak, "Subspace distribution clustering hidden Markov model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 264–275, 2001.

[5] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1377–1390, 2007.

[6] X. Sun and Y. Zhao, "Integrate template matching and statistical modeling for speech recognition," in *Interspeech*, 2010.

[7] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, XA Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al., *The HTK book (version 3.4)*, Cambridge University Engineering Department, 2006.

[8] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[9] L. Jiang and X. Huang, "Subword-dependent speaker clustering for improved speech recognition," in *ISCSLP*, 2000.

[10] Y. Han, A. Hamalainen, and L. Boves, "Trajectory clustering of syllable-length acoustic models for continuous speech recognition," in *ICASSP*. IEEE, 2006, pp. 1169–1172.

[11] F. Korkmazskiy, B.-H. Juang, and F. Soong, "Generalized mixture of HMMs for continuous speech recognition," in *ICASSP*. IEEE, 1997, pp. 1443–1446.

[12] 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, no. 2, pp. 171, 1995.