UNSUPERVISED TRAINING OF SUBSPACE GAUSSIAN MIXTURE MODELS FOR CONVERSATIONAL TELEPHONE SPEECH RECOGNITION

Zejun Ma[1]*, Xiaorui Wang*[1]*, Bo Xu*[1,2]

Digital Content Technology Research Center[1], National Lab of Pattern Recognition[2],
Institute of Automation, Chinese Academy of Sciences, Beijing, 100190

{zjma,xrwang,xubo}@hitic.ia.ac.cn

ABSTRACT

This paper presents our preliminary works on exploring unsupervised training of subspace gaussian mixture models for under-resourced CTS recognition task. The subspace model yields better performance than conventional GMM model, particularly in small or middle-sized training set. As an effective way to save human efforts, unsupervised learning is often applied to automatically transcribe a large amount of speech archives. The additional auto-transcribed data may help to improve model accuracy. In this paper, experiments are carried out on two publicly available English conversational telephone speech corpora. Both GMM and SGMM model in combination with unsupervised learning are examined and compared in this paper.

Index Terms— Speech recognition with low resources, unsupervised learning, subspace acoustic model.

1. INTRODUCTION

Currently, the majority of state-of-art speech recognition systems relies on a large amount of transcribed speech data to robustly estimate HMM-GMM acoustic model. However, the acquisition of large training resources is a challenging task. Especially, the transcription of audio data typically involves expensive manual labors of language experts in particular, and is very time-consuming. In the case of under-resourced language or dialect in general, the collection of large training data is one of major bottlenecks for developing LVCSR system.

The unsupervised learning has been gaining popularity as a method to greatly reduce human efforts. Typical procedure of unsupervised learning involves using a seed model, trained on small or middle-sized hand-transcrbied corpus, to recognize a large amount of unlabeled speech data. The recognized hypotheses may be used together with manual transcriptions to re-train new acoustic model. This approach has been shown to be effective in Broadcast News and Broadcast Conversation recognition tasks[1, 2, 3, 4, 5, 6]. Another solution to the data sparsity is the recently proposed subspace gaussian mixture modeling approach. The SGMM model uses a set of relatively low-dimensioned vectors to capture variances between states' output probability distributions, while the majority of model parameters is shared across states. The more compact representation of SGMM model results in more robust estimation of parameters and improved performance than conventional GMM model, especially when the amount of available training data is limited[7, 8].

In our knowledge, this work may be the first attempt of applying unsupervised learning on subspace acoustic model. The additional improvement is expected by combining the above two methods and that is the motivation of our works. The key features of our work presented in this paper include: (1) the UBM, which is used to initialize main subspace model, is showed to provide more gains from bootstrapping on additional untranscribed speech data. (2) for recognition task in CTS data, the highly erroneous procedure of transcribing unlabeled speech data requires a more effective data filtering method. An improved lattice-based utterance confidence is proposed to enhance the reliability of automatic transcriptions.

The remainder of this paper is organized as follows. First, the definition of SGMM model is simply recapitulated in section 2. In section 3, the unsupervised training procedures of SGMM model is described in detail. A lattice-based data selection method is proposed to improve data filtering. Section 4 presents experimental setups and comparative results. Section 5 provides concluding remarks.

2. REVIEW OF SGMM MODEL

In Subspace Gaussian Mixture Model, the pdf (probability distribution function) emitted by HMM state j is modeled by a mixture of Gaussians:

$$p(x|j) = \sum_{i=1}^{I} w_{ji} \cdot \mathcal{N}(x; \mu_{ji}, \Sigma_i) \quad (1)$$

$$\mu_{ji} = \mathbf{M}_i \cdot \mathbf{v}_j \quad (2)$$

This work is supported by Tsinghua - Tencent Joint Laboratory for Internet Innovation Technology.

$$w_{ji} = \frac{\exp \mathbf{w}_i^T \cdot \mathbf{v}_j}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \cdot \mathbf{v}_j} \quad (3)$$

As shown in Eq.(2)-(3), the mean vector μ_{ji} and weight w_{ji} of each mixture are not direct model parameters. Instead, they are linearly or log-linearly dependent on vector $\mathbf{v}_j$. The $\mathbf{v}_j$ is a vector specific to state j and is used to model the variations between different HMM states. The mapping from state-specific vector $\mathbf{v}_j$ to mixture mean μ_{ji} and weight w_{ji} is through structure $\mathbf{M}_i$ and $\mathbf{w}_i$. The mean projection matrix $\mathbf{M}_i$ and weight projection vector $\mathbf{w}_i$, plus covariance matrix Σ_i, are globally shared among all states. The SGMM model has compact parametric representation because the dimension S of $\mathbf{v}_j$ is typically being around the same as the feature dimension D and often far less than the parameter size of each mixture in comparison with diagonal covariance GMM, i.e. $S \approx D$ and $S \ll I \times (2D+1)$. Thus, the $\mathbf{v}_j$ lies in a subspace of GMM parameter space. In addition, a SGMM state can be extended to contain substates and this can be viewed as a tradeoff between model complexity and accuracy, like shown in Eq.(4)-(6). The speaker subspace and CMLLR transformation is not considered in this paper.

$$p(x|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \cdot \mathcal{N}(x; \mu_{jmi}, \Sigma_i) \quad (4)$$

$$\mu_{jmi} = \mathbf{M}_i \cdot \mathbf{v}_{jm} \quad (5)$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \cdot \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \cdot \mathbf{v}_{jm}} \quad (6)$$

3. UNSUPERVISED LEARNING OF SGMM

3.1. General procedure of unsupervised training

The procedure of unsupervised learning used in this paper is similar to [5, 9] and is applicable for both GMM and SGMM models. It consists of following steps:

- **Seed model training:** initial model (seed model) is trained on small amounts of manually transcribed speech data.
- **Transcription generation:** the seed model is used to recognize a corpus of unlabeled speech segments. The recognized 1-best hypotheses or word lattices are used together with manually transcribed corpus for latter model retraining.
- **Data selection:** the confidence score is used to measure the reliability of recognized transcriptions. The segments with confidences below a predefined threshold are simply omitted. A lattice-based utterance confidence measure is proposed in this paper to enhance data filtering. The detailed description of our data selection method is presented in subsection 3.3.
- **Model retraining:** the new acoustic model can be trained on the enlarged transcribed speech corpus.

The above procedure may be applied iteratively and the acoustic model can be incrementally refined. The criterion used to estimate acoustic model parameter in this study is restricted to the *Maximum Likelihood* framework, but the work can be extended to discriminative training[2].

3.2. Bootstrapping UBM

Although the Universal Background Model (UBM) doesn't appear in SGMM model definition, however it serves two important purposes: (1) a prototype to initialize SGMM model; (2) pruning Gaussians, i.e. Gaussian-selection, during likelihood computation. Essentially, the UBM is a generic mixture of Gaussians that models all speech classes (phonemes and silence)[7] and thus it can be trained without transcriptions. Usually, the UBM training is a separate process from the training of SGMM. Based on the above facts, the unlabeled data may be used together with hand-transcribed data to refine UBM model and the scheme is called "UBM bootstrapping" in our work.

3.3. Data selection

Usually, 1-best recognized hypothesis is used as a good approximation of the true transcription in unsupervised learning. The unique hypothesis is then used to align between audio data and phone models for EM-based HMM training. Alternative strategy is using lattice, instead of 1-best hypothesis, to represent recognition result. The lattice can be viewed as a compact representation of multiple possible hypotheses. These hypotheses in lattice may be applied to alleviate the adverse effect of recognition errors[1]. Moreover, the lattice can be also used to compute posterior probability of hypothesized word or whole utterance. The posterior probability is often applied as an indicator for quality of recognized hypotheses.

In general, the posterior probability of word W which occurs in $[t_s, t_e]$ can be computed by following equation:

$$\mathrm{PP}_{lat} = \frac{\sum_{W_-, W_+} \mathrm{P}(O t_s t_e | W_- W W_+) \mathrm{P}(W_- W W_+)}{\sum_{W'} \mathrm{P}(O|W') \mathrm{P}(W')} \quad (7)$$

with $\mathrm{PP}_{lat}(W t_s t_e | O)$ is the sum of the probabilities of all paths that contain the hypothesis word W from t_s to t_e, W_- and W_+ denoting any word sequence before t_s and after t_e respectively, W' being any word sequence. Eq.(7) can be efficiently implemented by the well-known forward-backward algorithm. Similarly, the posterior probability of most probable hypothesis $\hat{\pi}$ can be obtained with the numerator replaced by the likelihood of $\hat{\pi}$ in Eq.(7). The posterior probability of $\hat{\pi}$ is further normalized by the number of hypothesized words N and utterance duration L of $\hat{\pi}$. As scaling factor γ increases

the result with more hypothesized words and longer duration is more preferred. The parameter γ is empirically determined on the development set .

$$\mathrm{CM}_{lat}(\hat{\pi}) = \mathrm{PP}(\hat{\pi})^{\frac{\gamma}{N \cdot L}} \quad (8)$$

4. EXPERIMENTAL SETUPS AND RESULTS

The experiments in this paper are carried out using HTK, Kaldi speech recognition toolkit and SRILM toolkit to develop GMM, SGMM acoustic model and N-Gram backoff language model for the English system and tested for the CTS task.

4.1. Corpora

The manually transcribed corpus for acoustic model training is Callhome English CTS trainig set released by LDC in 1997. The amount of available data in this training set is about 15 hours. The assuming untranscribed corpus used in this work contains total 30 hours speech data collected from three sources: 11h from SWB1, 13h from Fisher1 and 6h from Fisher2.

Two test sets are used to evaluate the systems, 2h Callhome English evaluation set `CH97` and 8h RT03 evaluation set `RT03`.

4.2. System description

In signal processing module, the analysis frame length and shift are 25ms and 10ms respectively. The speech frames are parameterized as PLP (Perceptual Linear Prediction) features. The 39 dimensional feature includes 12 PLP coefficients plus energy with their first-order and second-order derivatives. The cepstral mean and variance normalization are also applied.

The acoustic model is a context-dependent HMM model with 1816 tied states obtained from decision tree clustering. Each triphone model has the left-to-right topology with 3 emitting states and the pdf (probability distribution function) of each state consists of 16 diagonal-covariance Gaussian mixtures. In SGMM modeling, there are 400 full-covariance Gaussian mixtures per state and the subspace dimension being equal to 40.

The bigram language model is interpolated using English Callhome, Switchboard1 and Fisher2 training corpus and it contains 35K words. The corpus for language model training is manually checked to ensure that it contains no transcriptions of assumed untranscribed corpus. The perplexity of language model on Callhome English evaluation set is 156.89, and 189.21 on RT03 evaluation set respectively. During all experiments, the language model is not changed since the main concern is to study the unsupervised training of acoustic model.

4.3. Results

4.3.1. baseline and UBM bootstrapping

The baseline acoustic models are trained on 15 hours of manually transcribed Callhome training data and evaluated on two test sets. The first entry in Table 1 gives the WER performance of GMM model, while the rest two entries for SGMM models. Since the baseline models are trained only on Callhome data, the wer on `CH97` are much lower than wer on `RT03`. The second and third entry in Table 1 represent baseline SGMM models initialized by different UBMs: the UBM in **SGMM** is trained only on 15h hand-transcribed data while in **SGMM-UBM**, total 45h data, i.e. 15h hand-transcribed data plus additional 30h untranscribed data, are used together to enhance UBM. It can be observed that there is benefit from using bootstrapped UBM to initialize main subspace model. Thus, the subspace models in **SGMM-UBM** are used as the baseline models for the rest of this paper.

Table 1. The WER(%) performance of baseline seed model.

Model	#substates	CH97	RT03
GMM	1816	55.78	64.54
SGMM	1816	51.75	61.95
	4000	50.44	60.76
	6000	50.00	60.51
SGMM-UBM	1816	51.05	61.55
	4000	50.22	60.43
	6000	49.78	60.00

4.3.2. data selection

Before carrying out unsupervised acoustic model training, the impact of different confidence measures on data selection is assessed. The baseline SGMM with 6K substates is used to decode 30h untranscribed data. Table 2 shows the WER of two types of lattice-based confidences with threshold value varying from 0.5 to 0.9. The WER is obtained by comparing 1-best hypotheses with corresponding true transcriptions. It can be seen that with the extreme values of threshold 0.0, meaning no data to be filtered, the poor WER 55.44% has necessitated the data selection. On the other hand, with tight threshold 0.9, more data were removed and only a small portion of data remained. Clearly, the threshold serves as a way to trade off between the size and quality of usable data. The columns with labels **word-conf** and **utter-conf** in Table 2 represent the effect of word and utterance confidence in data selection, respectively. It can be observed that there are substantial improvements of quality for remained usable data using utterance confidence over the conventional word confidence. Thus, the utterance confidence is adopted in the rest of this paper for data filtering.

Table 2. The comparison of WER on untranscribed data using two confidence measures.

threshold	word-conf		utter-conf	
	data size	wer(%)	data size	wer(%)
0.5	23h	42.31	21h	38.11
0.6	19h	40.13	17h	35.60
0.7	15h	34.56	13h	31.84
0.8	11h	30.87	10h	26.44
0.9	9h	25.69	7h	20.08
0.0	data size: 30h, wer: 55.44%			

4.3.3. unsupervised training

After data selection available auto-transcribed data are generated as described in the previous section. Table 3 gives the WER of both SGMM and GMM models trained on enlarged training data obtained in unsupervised manner. The impact of confidence threshold, ranging from 0.5 to 0.9, on performance is also examined in Table 3. The iterative addition scheme of auto-transcribed data is not considered in our work since the total amount of additional data is relatively small with respect to the hand-transcribed training data, 30h v.s. 15h. Thus, the additional data are directly used together with hand-transcribed data as re-training data. It is observed that the threshold 0.6 provides the best performance for both SGMM and GMM system. The relative WER reductions of SGMM with 6K substates over its baseline are 3.1% on `CH97` and 3.0% on `RT03`; the gains for GMM case is limited: 1.5% on `CH97` and 1.9% on `RT03`.

Table 3. The effect of confidence threshold on performance of unsupervised learning.

CH97 evaluation set						
model	#substates	WER(%)				
		0.9	0.8	0.7	0.6	0.5
SGMM	1816	50.57	50.42	50.31	50.27	50.29
	4000	49.92	49.65	49.41	49.35	49.41
	6000	49.33	49.12	48.62	**48.24**	48.87
GMM	1816	55.32	55.13	55.03	**54.92**	54.99
RT03 evaluation set						
model	#substates	WER(%)				
		0.9	0.8	0.7	0.6	0.5
SGMM	1816	60.60	60.23	59.60	59.50	59.53
	4000	60.04	59.65	58.97	58.99	59.12
	6000	59.90	59.12	58.67	**58.17**	58.60
GMM	1816	64.23	64.02	63.84	**63.26**	63.51

5. CONCLUSIONS

In this paper, we have combined compact SGMM modeling with unsupervised learning for under-resourced CTS recognition task. The UBM bootstrapping makes full use of untranscribed data to provide a better initialization for main subspace model; the utterance confidence is proposed to improve data filtering. Experimental results demonstrate that using the suggested training procedure the manual efforts of transcribing speech data can be greatly reduced for low-resourced scenarios.

6. REFERENCES

[1] T. F. Silva, J. L. Gauvain, and L. Lamel, "Lattice-based unsupervised acoustic model training," in *ICASSP*, Prague, Czech, May 2011, pp. 4656–4659.

[2] L. Wang, M. J. F. Gales, and P. C. Woodland, "Unsupervised training for mandarin broadcast news and conversation transcriptions," in *ICASSP*, Honolulu, Hawaii, April 2007, vol. IV, pp. 353–356.

[3] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: recent experiments," in *Eurospeech*, Budapest, Hungary, September 1999, pp. 2725–2728.

[4] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *ICASSP*, Toulouse, France, May 2006, vol. III, pp. 1056–1059.

[5] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," in *Automatic Speech Recognition and Understanding Workshop*, Trento, Italy, December 2001.

[6] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, February 1998, vol. IV, pp. 301–305.

[7] D. Povey, L. Burget, and M. Agarwal et al., "The subspace gaussian mixture model - a structured model for speech recognition," *Computer Speech and Language*, pp. 404–439, 2011.

[8] D. Povey, L. Burget, M. Agarwal, and P. Akyazi et al., "Subspace gaussian mixture models for speech recognition," in *ICASSP*, Dallas, Texas, USA, March 2010, pp. 4330–4333.

[9] C. Gollan, S. Hahn, R. Schlueter, and H. Ney, "An improved method for unsupervised training of lvcsr system," in *Interspeech*, Antwerp, Belgium, August 2007, pp. 2101–2104.