FEATURE AND SCORE LEVEL COMBINATION OF SUBSPACE GAUSSINAS IN LVCSR TASK

Petr Motlicek[1*], *Daniel Povey*[2], *Martin Karafiat*[3]

[1] Idiap Research Institute, Martigny, Switzerland
[2] Johns Hopkins University, Baltimore, USA
[3] Brno University of Technology, Czech Republic
motlicek@idiap.ch, dpovey@gmail.com, karafiat@fit.vutbr.cz

ABSTRACT

In this paper, we investigate employment of discriminatively trained acoustic features modeled by Subspace Gaussian Mixture Models (SGMMs) for Rich Transcription meeting recognition. More specifically, first, we focus on exploiting various types of complex features estimated using neural network combined with conventional cepstral features and modeled by standard HMM/GMMs and SGMMs. Then, outputs (word sequences) from individual recognizers trained using different features are also combined on a score-level using ROVER for the both acoustic modeling techniques. Experimental results indicate three important findings: (1) SGMMs consistently outperform HMM/GMMs (relative improvement on average by about 6% in terms of WER) when both techniques are exploited on single features; (2) SGMMs benefit much less from feature-level combination (1% relative improvement) as opposed to HMM/GMMs (4% relative improvement) which can eventually match the performance of SGMMs; (3) SGMMs can be significantly improved when individual systems are combined on a score-level. This suggests that the SGMM systems provide complementary recognition outputs. Overall relative improvements of the combined SGMM and HMM/GMM systems are 21% and 17% respectively compared to a standard ASR baseline.

***Index Terms*—** Automatic Speech Recognition, Discriminative features, System combination

1. INTRODUCTION

Discriminative techniques for training probabilistic features used in Automatic Speech Recognition (ASR) have been extensively studied in the last decade. The first probabilistic features exploited in Gaussian Mixture Model (GMM) based HMMs have been proposed in [1]. Phone posterior probability estimates obtained from discriminatively trained artificial Neural Network (NN) and then post-processed were used as inputs for HMM/GMMs. Although preliminary versions of such the NN based features did not outperform conventional cepstral features, interestingly, they have shown complementary performance and thus their subsequent combination on a feature-level brought significant ASR improvements. Recently, more complex NN based features have been proposed using a Bottle-Neck (BN) approach [4, 5]. Although the features are also obtained as a product of NNs, they are not derived from the phone-class posteriors. Instead, the features are obtained as linear outputs from a middle (bottle-neck) layer in a 5-layer NN. Nowadays, BN features (combined with conventional MFCCs [2] or PLPs [3]) modeled using HMM/GMMs constitute a state-of-the-art in Large Vocabulary Continuous Speech Recognition (LVCSR) task [5, 6, 7].

In acoustic modeling, a significant effort has been directed in last years toward model adaptation and multilingual approaches. Among others, Subspace Gaussian Mixture Models (SGMMs) have been proposed [8]. Unlike conventional HMM/GMMs, SGMMs split the model into globally shared parameters and parameters specific to acoustic states which enables various kinds of acoustic model tying. Such the new model structure has been successfully explored in the multilingual acoustic model adaptation [9]. Besides the model adaptation tasks, SGMMs have also been explored in monolingual ASR tasks, especially in constrained recognition scenarios (e.g., read speech, small-vocabulary tasks) [10, 11], but preliminary evaluations were also performed on an LVCSR scenario [12].

In this paper, we investigate employment of state-of-the-art BN features and their combination with conventional cepstral features in an SGMM framework. Since experimental results indicate that SGMMs do not benefit from the feature-level combination, as opposed to HMM/GMMs, we analyze complementarity of individual systems for both acoustic modeling techniques. To estimate a measure of complementarity, we use ROVER - Recognizer Output Voting Error Reduction - a technique allowing to combine word (symbol) sequences taken as outputs of different recognition systems [13]. In our experiments, neural networks and acoustic models are trained on 150 hours of meeting data and evaluated on well-known NIST Rich Transcription (RT'07) ASR evaluation task[1]. We demonstrate that although SGMMs do not benefit from feature-level combination, significant improvements can be achieved by combining recognition outputs on a score-level. Eventually, amount of parameters to be estimated for the SGMM systems are considerably less than for the HMM/GMM systems.

In the reminder of this paper, we first review the concept of NN based features as well as subspace GMMs (Section 2). Then, Section 3 introduces our experimental setup and used datasets. Experimental results on feature-level and score level combinations are presented in Sections 4 and 5, respectively. Section 6 concludes the work.

[*] This work was supported by the Swiss Commission for Technology and Innovation under the project name "TAO-CSR: Task Adaptation and Optimisation for Conversational Speech Recognition" and by Eurostars Programme powered by Eureka and the European Community under the project "D-Box: A generic dialog box for multilingual conversational applications". P. Motlicek was supported by Samsung Electronics Co. Ltd, South Korea, under the project "Multi-Lingual and Cross-Lingual Adaptation for Automatic Speech Recognition". M. Karafiat was supported by Grant Agency of the Czech Republic post-doctoral project No. P202/12/P604.

[1] http://www.itl.nist.gov/iad/mig/tests/rt/2007/index.html

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

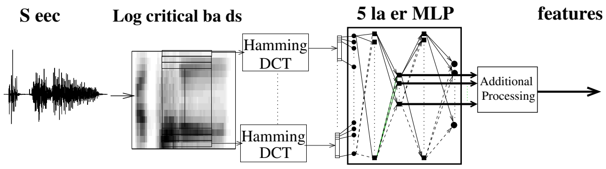

Fig. 1. Diagram of extracting BN features using 5-layer NN.

2. RELATED WORK

In this section, we first review the concept of NN based features and then briefly summarize the SGMM acoustic modeling framework.

2.1. NN based features

The probabilistic features are usually considered as phone class posterior probabilities given the acoustics and estimated with a NN that can be trained on any auxiliary dataset The language of the training data determines the number of output units K (number of phone classes) of the NN. The phone classes can for example be context-independent monophones or context-dependent triphones.

Unlike phone posteriors estimated using traditional 3-layer NN, we exploit Bottle-Neck (BN) features obtained from a 5-layer NN where the middle hidden layer (BN layer) has the size of the desired feature vector. The first and the third hidden layers in the NN are usually of the same size. The choice of using 5-layer NN is satisfied by their significantly higher performance achieved already on a frame-level during NN training, as shown later in Section 4. As illustrated in Fig. 1, the NN is trained using spectral based features extended by a temporal context [5]. First, the critical-band-energies are extracted from the speech. A 23 Mel-scaled filter bank is used for 16 kHz signal. Further, a block of consecutive 31 frames is created representing a 310 ms long temporal context and each energy coefficient is post-processed by applying Hamming window and Discrete Cosine Transform (DCT). Eventually, the first 16 DCT coefficients are preserved in each critical-band and concatenated (over all 23 spectral bands) into the final 368 (16×23) dimensional feature vectors used for NN training. Unlike conventional Tandem approach where features are represented by phone-posterior estimates, the BN features generated usually using a 5-layer NN are obtained as linear outputs in the third (bottle-neck) NN layer.

2.2. SGMMs

Subspace Gaussian Mixture Models (SGMMs) enable to compactly represent a large collection of mixture-of-Gaussian models. Unlike conventional HMM/GMMs in which state model parameters are directly estimated from the data, SGMM model parameters are derived from a set of state specific parameters, and from a set of globally shared parameters which can capture phonetic and speaker variation [8].

In the case of a conventional GMM, the likelihood is given as

$$p(\mathbf{x} \mid j) \quad = \quad \sum_{i=1}^{M_j} w_{ji}\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji}), \tag{1}$$

where j is the state and the parameters of the model are w_{ji}, $\boldsymbol{\mu}_{ji}$ and $\boldsymbol{\Sigma}_{ji}$. The SGMM in the basic case is given as

$$\begin{aligned} p(\mathbf{x} \mid j) &= \sum_{i=1}^{I} w_{ji}\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{i}) && (2) \\ \boldsymbol{\mu}_{ji} &= \mathbf{M}_i \mathbf{v}_j && (3) \\ w_{ji} &= \frac{\exp \mathbf{w}_i^T \mathbf{v}_j}{\sum_{l=1}^{I} \exp \mathbf{w}_l^T \mathbf{v}_j}, && (4) \end{aligned}$$

where $\mathbf{v}_j$ are state specific vectors (with dimension similar to that of the speech features), and $\mathbf{w}_i$, $\mathbf{M}_i$, and $\boldsymbol{\Sigma}_i$ are globally shared parameters. I is the number of Gaussians in the shared GMM structure. In fact, we employ a Universal Background Model (UBM) which is a mixture of full-covariance Gaussians of size I that is used to initialize the system and to prune the Gaussian indices during training and decoding. The basic concept of SGMMs can be extended towards large-scale acoustic models by adding sub-state specific vectors and speaker-dependent mean offsets via speaker vector parameters $\mathbf{v}^{(s)}$ and "speaker projections" $\mathbf{N}_i$ [12]. Sub-state specific vectors represent a way of largely extending the model capacity while preserving the total number of parameters. In the following Sections 4 and 5, SGMMs will also be extended with speaker vectors towards a speaker-dependent system to demonstrate their efficiency for speaker-dependent acoustic modeling.

NN structure	Train	CV
3-layer	67.3	66.4
3-layer (spk norm)	68.5	67.5
5-layer (spk norm)	70.9	69.7

Table 1. Acc [%]: Frame-based phone accuracies estimated for training and Cross-Validation (CV) sets for different NN structures.

3. EXPERIMENTAL SETUP

All the experiments were done with the open-source Kaldi speech recognition toolkit [11]. Our LVCSR system is partially following the AMI-LVCSR system represented by quite a complex approach running in several passes and developed for NIST RT'07 (meeting data) evaluations [14].

For detailed analysis of acoustic modeling techniques, only one-pass ASR system is implemented. Instead of applying VTLN, CMLLR and expanding lattices using four-gram Language Model (LM), one-pass decoding is performed using a bi-gram LM. The dictionary contains around 50 K words. The acoustic scale factors were always tuned for the best Word-Error-Rates (WERs) during our experiments. AMI and ICSI meeting data yielding in total 150 hours of segmented speech is exploited for training of NNs and acoustic models. The data is represented by Individual Head Microphone (IHM) recordings sampled at 16 kHz and the reference segmentation is used.

As a baseline, conventional mean- and variance-normalized (per speaker) MFCCs and PLPs expanded using their first and second order derivatives (39 dimensions) are initially evaluated using HMM/GMMs. Similar to [7], we also exploit 3rd order derivatives in PLPs subsequently reduced by HLDA (described in [15]) to 39 dimensional features. The HLDA considers each Gaussian component as a class. Cross-word tied-states triphone HMM/GMMs (having diagonal covariance matrices) were trained by Maximum Likelihood (ML). The model contains 5 K tied states and in total 120 K Gaussians. Performance of the baseline systems is given in Tab. 2. Com-

features	dimension	HMM/GMM		SGMM		SGMM+“spkvecs”	
		WER [%]	# params	WER [%]	# params	WER [%]	# params
MFCC	39	42.1	(9.4M)	37.9	(6.4M)	36.2	(7.1M)
PLP	39	41.7		38.4		36.6	
PLP$_{HLDA}$	39	39.2		38.9		36.5	
BN	30	37.6	(7.2M)	35.9	(6.0M)	35.0	(6.5M)
BN$_{PCA}$	30	37.5		35.8		35.0	
BN$_{HLDA}$	30	37.1		36.1		34.7	
BN$_{HLDA+\Delta}$	60	35.8		35.2		34.5	
PLP+BN	69	35.9	(16.6M)	**34.6**	(7.9M)	**34.0**	(9.3M)
PLP+BN$_{HLDA}$	69	35.3		34.9		34.6	
PLP$_{HLDA}$+BN	69	35.4		34.9		34.2	
PLP$_{HLDA}$+BN$_{HLDA}$	69	35.0		34.8		34.2	
PLP$_{HLDA}$+BN$_{HLDA+\Delta}$	99	**34.6**	(23.8M)	34.7	(9.9M)	34.1	(11.8M)

Table 2. WER[%]: Performance of different features and their combinations modeled by HMM/GMMs and SGMMs. We also estimate amount of parameters of the corresponding acoustic models. Bold numbers highlight the best systems.

pared to [7], slightly less training data (150 hours instead of 180 hours) was used without VTLN normalization.

In the following experiments using HMM/GMMs, we apply a concept of Single Pass Retraining (SPR) where an initial model was always trained on simple PLPs. Our informal experimental results indicate that SPR when exploited on BN features performs similar to the full GMM training. The same HMM/GMM model size is therefore kept after the SPR. Eventually, 12 ML iterations are followed to better settle new GMMs in the new feature space.

4. FEATURE-LEVEL COMBINATION

First, we describe extraction of “simple” BN features used throughout our experiments. For NN, 124 hours of randomly selected data from AMI/ICSI corpus was used for training and 12 hours for Cross-Validation (CV). We decided to use a 5-layer NN topology as it was shown to outperform 4-layer NN [6]. Inspired by [7], the final size of 5-layer NN was selected to have about 2M parameters for 368 dimensional input vectors (per speaker mean- and variance-normalized), for NN trained to classify sub-phone classes (i.e., K=135 targets corresponding to 45 English phonemes uniformly split into 3-states). An increase in gain while exploiting subphone classes during training has been observed in [16]. In the case of probabilistic (Tandem) features, the gain which can be achieved from sub-phone classes goes at the expense of large dimensionality of the output features. However in the case of BN features, the number of output classes does not directly affect the output feature size and thus sub-phone classes can be easily used for the NN training. Based on our various informal experiments, the NN with bottle-neck size of 30 performed the best and the linear outputs were taken from the bottle-neck layer to create output features. For the selected NN-size, a possible 5-layer NN topology is 368-4 K-30-4 K-135. For the sake of comparison, Tab. 1 compares performance of a 5-layer NN (having 4 K neurons in hidden layers) with a conventional 3-layer NN alternative (having also 4 K neurons in the hidden layer). NN performance is presented for 1-state phone output and frame-based phone accuracies for the training and CV sets. The results clearly show that speaker normalization performed on top of input features and a 5-layer NN topology significantly improve discrimination of the NN.

Further, let us consider an HMM/GMM framework. Performances of BN and standard cepstral features for RT’07 ASR task are summarized in Tab. 2. BN features achieve expected WER improvements of about 4% absolute over PLPs. In addition, simple BN features were deccorelated using Principal Components Analysis (PCA) and also by previously mentioned HLDA prior HMM/GMM modeling. According to results presented in Tab. 2, HLDA is preferred over PCA. HLDA is assumed to maximize the between-class separability and in contrast to well-known Linear Discriminant Analysis (LDA), HLDA does not assume the class covariances to be the same. Then, inspired by [7], NN based features were extended with the first order derivatives (+Δ) which are expected to overcome an HMM assumption of frame-independence. This brings another 2% considerable improvement over simple BN features.

Eventually on the feature-level, we also evaluated a combination of cepstral and BN features (without any subsequent dimensionality reduction). Experimental results for various types of feature combinations, shown in Tab. 2, demonstrate that both PLPs and BN features are to some extent complementary and additional WER improvements (of about 1% absolute) can be achieved.

4.1. SGMMs

Similar to HMM/GMMs, SGMMs were first trained on standard cepstral features. For SGMMs, similar context-tree tying is exploited with 5 K states. The UBM is trained on the whole AMI/ICSI data and I=500 Gaussians are retained. The total-number of state-specific vectors is 100 K. Throughout all experiments, the subspace dimension was kept constant and equal to S=50 (in case of using speaker vectors, the dimension was kept equal to 39). Results in Tab. 2 clearly show that SGMMs significantly reduce WER (by about 3-4% absolute) compared to HMM/GMMs for standard cepstral features.

Then, BN features were explored. Unlike HMM/GMMs developed using SPR, SGMMs were always trained from scratch. As shown in Tab. 2, simple BN features applied in the SGMM framework reduce WER by about 4% compared to the cepstral features. Such the reduction is similar to the one achieved by HMM/GMMs. Interestingly, HLDA deccorelation applied prior SGMM modeling does not help. Similar trend can be observed for PLPs deccorelated by HLDA. Since UBM is trained to retain full-covariance GMMs, we hypothesize that this additional step of deccorelation is useless. Although NN is trained with large temporal context (i.e., 310 ms), an extension of the BN features by first order derivatives brings additional gain, similar to HMM/GMMs.

Finally, BN features were combined with PLPs. Once the best

PLP+MFCC	BN	BN$_{HLDA+\Delta}$
HMM/GMM	38.4	38.2
SGMM	34.8	34.6
SGMM+"spkvecs"	33.3	33.2

Table 3. WER[%]: Score-level combination of three individual systems using ROVER for three types of acoustic models: HMM/GMM, SGMM, SGMM+"spkvecs".

	PLP+MFCC+BN$_{HLDA+\Delta}$ + "BEST"
HMM/GMM	34.5
SGMM	32.9
SGMM+"spkvecs"	32.1

Table 4. WER[%]: Score-level combination of three individual systems plus the best system from Tab. 2 using ROVER for three types of acoustic models: HMM/GMM, SGMM, SGMM+"spkvecs".

feature-level combined systems are compared to the best individual systems for each acoustic modeling framework, we observe that SGMMs benefit much less from feature-level combination (marginal 1% relative improvement) as opposed to HMM/GMMs (about 4% relative). In addition as indicated in Tab. 2, HMM/GMMs eventually match the performance of SGMMs after performing the feature-level combination. Similar trends can be observed for speaker-dependent SGMMs (employing speaker vectors denoted to as "spkvecs") in Tab. 2.

4.2. Acoustic model size

In addition to WERs, Tab. 2 also shows total number of parameters of each individual acoustic model[2]. Although conventional HMM/GMMs perform similar to SGMMs after exploiting combined BN and PLP features (WER about 32.6%), SGMMs have about 3 times less parameters if the best performing systems are compared.

5. SCORE-LEVEL COMBINATION

Although SGMMs provide much better performance when employed over single acoustic features, feature-level combination produced marginal improvement. Such the trend is on the contrary to which was observed in the case of conventional HMM/GMMs. As a consequence, both acoustic modeling frameworks achieve similar performance (bold numbers in Tab. 2).

Unlike feature-level combination, this section focuses on combining individual recognition systems on a score-level. More partially, we employ ROVER - a standard technique allowing to combine word (symbol) sequences taken as outputs of different recognition systems [13]. ROVER can be seen as a simple approach measuring complementarity of recognition systems based on counting simultaneous and dependent errors. It assumes that significant recognition gain can be achieved if the combined systems exhibit different (heterogeneous) recognition errors.

Results on score-level combination for the both HMM/GMM and SGMM systems are given in Tab. 3. Outputs of three individual recognition systems are always combined (trained using PLPs, MFCCs and simple or HLDA-transformed BN features). Interestingly, a combination of HMM/GMM-based systems fails, since the ROVER output performs worse than the best (BN-based) individual system. However, SGMMs can well benefit from the score-level combination (WER=34.8% as opposed to the best (BN-based) individual system with WER=35.9%). This suggests that the SGMM-based recognizers trained with diverse features make heterogeneous errors at the output.

In addition to three individual recognition systems combined in Tab. 3, in Tab. 4, we use for the score-level combination also the "best" performing system (developed using feature-level combination based on results given in Tab. 2). Tab. 4 demonstrates the best final performance for the both HMM/GMMs as well as SGMMs acoustic modeling techniques (also a speaker-dependent SGMM framework is presented). Compared to the HMM/GMM-MFCC baseline, 17% and 21% relative improvements in WER are achieved for speaker-independent HMM/GMM and SGMM systems.

6. CONSLUSIONS

We have demonstrated that the SGMM framework is an efficient approach in the LVCSR task. Overall evaluations of SGMMs exploiting powerful but complex PLP-BN features yield similar results as those obtained by conventional HMM/GMMs. Nevertheless, the total number of SGMM parameters is about 3 times less than in the HMM/GMM framework. Evaluation results also indicate different properties of the examined acoustic modeling techniques. Although SGMMs consistently outperform HMM/GMMs when built over individual features, HMM/GMMs can benefit much more from the feature-level combination than SGMMs. Nevertheless based on an analysis measuring complementarity of individual recognition systems, we show that SGMM-based recognizers produce heterogeneous outputs (scores) and thus subsequent score-level combination can bring additional improvement.

7. REFERENCES

[1] H. Hermansky, D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," *in Proc. of ICASSP*, pp. 1635-1638, Turkey, 2000.

[2] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *in IEEE Trans. on ASSP*, 28(4):357-366, 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *in The Journal of the Acoustical Society of America*, vol. 87, pp. 1738, 1990.

[4] S. Dupont, C. Ris, O. Deroo and S. Poitoux, "Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents," *in Proc. of ASRU*, pp. 29-34, Mexico, November 2005.

[5] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," *in Proc. of ICASSP*, pp. 757-760, Honolulu, USA, 2007.

[6] F. Grezl and P. Fousek, "Optimizing bottle-neck features for LVCSR," *in Proc. of ICASSP*, pp. 4729-4732, Las Vegas, USA, 2008.

[7] F. Grezl, M. Karafiat and L. Burget, "Investigation into bottle-neck features for meeting speech recognition," *in Proc. of Interspeech*, pp. 2947-2950, Brighton, GB, 2009.

[2] Note: amount of parameters of the NN is not included.

[8] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. karafiat, A. Rastrow, R. C. Rose, P. Schwarz and S. Thomas, “The Subspace Gaussian mixture model - A structured model for speech recognition,” In Computer Speech & Language, vol. 25, no. 2, pp. 404-439, 2011.

[9] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. karafiat, D. Povey, A. Rastrow, R. C. Rose and S. Thomas, “Multilingual Acoustic Modeling For Speech Recognition Based On Subspace Gaussian Mixture Models”, *in Proc. of ICASSP*, pp. 4334-4337, Dallas, USA, 2010.

[10] T. N. Vu, T. Schultz and D. Povey., “Modeling Gender Dependency in the Subspace GMM Framework,” *in Proc. of ICASSP*, pp. 4345-4348, Japan, 2012.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer and K. Vesely, “The Kaldi Speech Recognition Toolkit,” *in Proc. of ASRU*, Hawai, USA, December 2011.

[12] D. Povey, M. Karafiat, A. Ghoshal and P. Schwarz, “A Symmetrization of the Subspace Gaussian Mixture Model”, *in Proc. of ICASSP*, pp. 4504-4507, Prague, Czech Republic, 2011.

[13] J. G. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” *in Proc. of ASRU*, pp. 347-354, USA, 1997.

[14] T. Hain, L. Burget, J. Dines, G. Garau, V. Wan, M. Karafiat, J. Vepa and M. Lincoln, “The AMI system for the transcription of speech meetings,” *in Proc. of ICASSP*, pp. 357-360, Honolulu, USA, April 2007.

[15] N. Kumar and A. G. Andreou, “Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition,” *Speech Communication*, 26(4):283-297, 1998.

[16] P. Schwarz, P. Matejka and J. Cernocky, “Towards lower error rates in phoneme recognition,” *in Lecture Notes in Computer Science*, pp. 465-472, Brno, 2004.