View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

IDIAP RESEARCH REPORT

APPLICATION OF SUBSPACE GAUSSIAN MIXTURE MODELS IN CONTRASTIVE ACOUSTIC SCENARIOS

Petr Motlicek Philip N. Garner David Imseng
Fabio Valente

Idiap-RR-20-2012

Version of DECEMBER 10, 2012

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Application of Subspace Gaussian Mixture Models in Contrastive Acoustic Scenarios

Petr Motlicek[1], *Philip N. Garner*[1], *David Imseng*[1,2], *Fabio Valente*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{motlicek,pgarner,dimseng,valente}@idiap.ch

Abstract

This paper describes experimental results of applying Subspace Gaussian Mixture Models (SGMMs) in two completely diverse acoustic scenarios: (a) for Large Vocabulary Continuous Speech Recognition (LVCSR) task over (well-resourced) English meeting data and, (b) for acoustic modeling of under-resourced Afrikaans telephone data. In both cases, the performance of SGMM models is compared with a conventional context-dependent HMM/GMM approach exploiting the same kind of information available during the training. LVCSR systems are evaluated on standard NIST Rich Transcription dataset. For under-resourced Afrikaans, SGMM and HMM/GMM acoustic systems are additionally compared to KL-HMM and multilingual Tandem techniques boosted using supplemental out-of-domain data. Experimental results clearly show that the SGMM approach (having considerably less model parameters) outperforms conventional HMM/GMM system in both scenarios and for all examined training conditions. In case of under-resourced scenario, the SGMM trained only using in-domain data is superior to other tested approaches boosted by data from other domain.

Index Terms: large vocabulary continuous speech recognition, acoustic modeling, under-resourced languages

1. Introduction

Conventional acoustic modeling technique in Automatic Speech Recognition (ASR) represents distributions of (usually tied) Hidden Markov Model (HMM) states using relatively large number of parameters completely defining a Gaussian Mixture Model (GMM). This approach nowadays constitutes the state-of-the-art in acoustic modeling for ASR. This is especially valid for Large Vocabulary Continuous Speech Recognition (LVCSR). The main advantage of the HMM/GMM compared to other acoustic modeling techniques is its feasibility for parallel training (i.e., it can easily accommodate large amount of training data which is usually available for well-resourced languages) and possibility to combine standard adaptation and discriminative training techniques.

In case of under-resourced corpora (i.e., less "vivant" languages), and/or small vocabulary ASR systems, many new techniques have revealed, such as Kullback-Leibler divergence based Hidden Markov Models (KL-HMM) or multilingual Tandem systems [1]. Similar to HMM/GMM, these approaches can directly be trained with the data of an under-resourced corpora. In addition, they show large benefit over the conventional context-dependent HMM/GMM if they are boosted by properly combining acoustic information from multiple (e.g., out-of-domain, different languages) corpora [2].

Recently, a new acoustic modeling scheme based on Subspace Gaussian Mixture Model (SGMM) has been proposed [3]. Similar to KL-HMM or multilingual Tandem systems, SGMMs demonstrated their large potential to benefit from available data from different corpora (i.e., well-resourced languages) to improve recognition performance of the target domain (i.e., under-resourced language) [4]. Compared to other (multilingual) techniques, such as traditional ones exploiting universal phone models to allow for training acoustic models from many languages [5], SGMM (as well as KL-HMM and multilingual Tandem) can utilize a target phone set thus representing much simpler procedure.

To our knowledge, the SGMM, if trained for acoustic modeling by exploiting purely in-domain data, was evaluated only on standard (less ambitious) datasets (Wall-Street Journal, Resource Management tasks) [6]. In this paper, we evaluate the SGMM technique for acoustic modeling on two challenging but completely diverse speech recognition tasks: (a) LVCSR task performed on 16 kHz meeting data (well-resourced English language utilizing 150 hours of training data), and evaluated over standard NIST Rich Transcription (RT) 2007 data and, (2) small-vocabulary ASR over 8 kHz telephone speech (under-resourced Afrikaans utilizing only 3 hours of training data). Experiments reveal that the SGMM approach is superior to the conventional HMM/GMM technique on the LVCSR task for all examined training conditions. Furthermore, although the SGMM did not profit in our studies from additional source of data, as this was shown to be possible according to the multilingual experiments described in [4], the results on Afrikaans (considered as an under-resourced dataset) demonstrate significantly better performance compared to KL-HMM and multilingual Tandem systems boosted by out-of-domain data. Most of our experimental acoustic systems were built using Kaldi toolkit[1].

The paper is organized as follows: In Section 2, we give an overview of the Subspace Gaussian Mixture Model and its relation to the conventional HMM/GMM approach. Section 3 describes the systems used in our experimental work and Section 4 presents experimental results. Section 5 concludes the work.

This work was partially supported by Samsung Electronics Co. Ltd, South Korea under the project name "Domain Adaptation Using Subspace Model"; by the Swiss National Center of Competence in Research (NCCR) on "Interactive Multi-modal Information Management (IM)2"; and by the Swiss Commission for Technology and Innovation under the project name "TAO-CSR Task Adaptation and Optimisation for Conversational Speech Recognition".

[1]http://kaldi.sourceforge.net

2. Subspace Gaussian Mixture Models

The main goal of SGMM models is to reduce number of parameters by selecting the Gaussians from a subspace spanned by a Universal Background Model (UBM) of size I and state-specific transformations. The emission probability density functions of the SGMM can be estimated as [3]:

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

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i\mathbf{v}_j \quad (2)$$

$$w_{ji} = \frac{\exp \mathbf{w}_i^T\mathbf{v}_j}{\sum_{l=1}^{I}\exp \mathbf{w}_l^T\mathbf{v}_j}, \quad (3)$$

where $\mathbf{x} \in R^D$ denotes feature vector, j is the speech state, and $\mathbf{v}_j \in R^S$ is the state-specific vector (of dimension S). The model in each HMM state is represented by a simple GMM with I Gaussians, mixture weights w_{ji}, means $\boldsymbol{\mu}_{ji}$, and covariances $\boldsymbol{\Sigma}_i$ which are shared among the states. The state-specific vectors $\mathbf{v}_j \in R^S$ of the "subspace dimension" S (where S is typically around the same as dimensionality of acoustic features) together with globally shared parameters $\mathbf{M}_i$ and $\mathbf{w}_i$ are used to derive the means and mixture weights representing the given HMM state.

In fact, the previous set of equations assume one state-specific vector $\mathbf{v}$ to be assigned to each HMM-state. In order to allow for more precise modeling of HMM states, each state is rather represented with mixture of sub-states [7].

2.1. Additional speaker vectors

An useful extension to the basic SGMM framework is provided by using speaker vectors, which can be very beneficial especially in the LVCSR tasks. In this extension, as proposed in [7], each speaker s is described by a speaker vector $\mathbf{v}^{(s)}$ of an arbitrary dimension T (usually $T \sim S$). The mean projection (previously given by Equation 2) will now become:

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i\mathbf{v}_j + \mathbf{N}_i\mathbf{v}^{(s)}, \quad (4)$$

where the $\mathbf{N}_i$ matrices define the speaker subspace.

3. System descriptions

3.1. Well-resourced LVCSR meeting system

We use Individual Head Microphone (IHM) recordings sampled at 16 kHz for the LVCSR experiments with well-resourced data. The training set consists of the complete AMI and ICSI meeting data yielding a total of 150 hours of the segmented speech. The test set is defined by NIST RT 2007 evaluations[2]. The IHM condition of the "conference room meeting test set" with the reference segmentation was used in the experiments. The dictionary contains around 50k words and the decoding is performed using bi-gram Language Model (LM). Two benchmark ASR systems were considered in this scenario, both trained using HTK toolkit[3] (the systems are marked with 'H').

3.1.1. HMM/GMM^H benchmark system using HLDA-PLPs

The performance of several various (usually complex multi-pass) LVCSR systems on NIST RT 2007 test sets can be found on the Web[2]. A one-pass HMM/GMM^H based LVCSR system trained with slightly more training data (180 hours) which exploits, apart from HLDA-PLP features, quite sophisticated Tandem features is presented in [8]. HMM/GMM^H models were trained in the Maximum-Likelihood framework. The final acoustic model contains around 5.6k tied states and employes 18 Gaussian mixture components per state. In our implementation, slightly less training data (around 4.5k tied states) without VTLN normalization was used.

3.1.2. HMM/GMM^H benchmark system using MFCCs

For the sake of comparison, we also implemented an HMM/GMM^H benchmark system using the HTK which is trained on the same amount of data as the following SGMMs. It employes standard per-speaker normalized PLP features accompanied by Δ and $\Delta\Delta$. We used around 4.5k tied states and 18 Gaussian mixture components per state.

3.2. Under-resourced small vocabulary ASR system

In the second experimental setup, we decided to use Afrikaans – a resource scarce language. Relatively small amount of data is available from LWAZI corpus provided by the Meraka Institute, South Africa [9]. In total, 3 hours of the training data and 50 minutes of the test data is available. The dictionary contains around 1.5k words [10]. Since we did not have access to an appropriate LM, a uni-gram word LM was trained on word transcriptions from the training set. Two benchmark ASR systems were also considered in this scenario, both mostly trained using HTK toolkit[3] (the systems are also marked with 'H').

3.2.1. KL-HMM^H and multilingual Tandem^H benchmark systems

To compare performance of the SGMM on Afrikaans data, we use KL-HMM^H and multilingual Tandem^H systems presented in [1], particularly developed to be applied in under-resourced scenarios. In [1], phoneme accuracies were reported. For the sake of comparison, we use the same uni-gram word LM, as for SGMM models, and report Word Error Rates (WER) for both KL-HMM^H and multilingual Tandem^H systems.

More specifically, both systems use phoneme posterior probabilities as features. To estimate phoneme posterior probabilities, four Multilayer Perceptrons (MLPs) were previously trained using PLP+Δ+$\Delta\Delta$ features: one MLP on Afrikaans data and three on out-of-language data (English and German SpeechDat(II)[4] and the Spoken Dutch Corpus [11]). All four MLPs were trained on context-independent phoneme targets. For the acoustic modeling only Afrikaans data was used, but processed by all four previously trained MLPs. First, context-independent monophone models were built and then used as seeds for the context-dependent phoneme model training. The resulting phoneme posterior features have a dimension of 189. For the multilingual Tandem^H system, we used 1.5k states, each modeled with a mixture of 8 Gaussians. Since the KL-HMM^H system models each state with one categorical distribution, we used larger number (15.5k) of states.

4. Experimental work and results

All the following ASR experiments are carried out using Kaldi toolkit (marked with 'K'). The standard features used

[2] http://www.itl.nist.gov/iad/mig/tests/rt/2007

[3] http://htk.eng.cam.ac.uk

[4] http://www.speechdat.org/SpeechDat.html

System	WER [%]	Description
Well-resourced LVCSR meeting system		
HMM/GMM^H	41.7	PLPs+Δ+$\Delta\Delta$
HMM/GMM^H	39.2	HLDA-PLPs
HMM/GMM^K	42.1	MFCCs +Δ+$\Delta\Delta$
HMM/GMM^K	39.4	MLLT-MFCCs
HMM/GMM^K	35.3	MLLT-MFCCs + SAT(fMLLR)
SGMM^K	38.2	MFCCs +Δ+$\Delta\Delta$
SGMM^K	36.2	MFCCs +Δ+$\Delta\Delta$ + speaker vectors
SGMM^K	34.4	MLLT-MFCCs + speaker vectors
SGMM^K	34.2	MLLT-MFCCs + speaker vectors + fMLLR

Table 1: *LVCSR experimental results of HMM/GMM and SGMM systems evaluated on NIST RT 2007 data. PLP and HLDA-PLP based HMM/GMM benchmark systems, implemented using HTK ('H'), are described in Section 3. All other systems, implemented using Kaldi ('K'), are presented in Section 4.*

System	N $\times 10^6$	Description
Well-resourced LVCSR meeting system		
HMM/GMM^H	6.4	PLPs+Δ+$\Delta\Delta$
HMM/GMM^H	6.4	HLDA-based PLPs
HMM/GMM^K	9.4	MFCCs +Δ+$\Delta\Delta$
SGMM^K	6.4	MFCCs +Δ+$\Delta\Delta$
Under-resourced small vocabulary ASR system on Afrikaans		
KL-HMM^H	3.0	Posterior features estimated on PLPs+Δ+$\Delta\Delta$
Tandem^H	4.5	Posterior features estimated on PLPs+Δ+$\Delta\Delta$
HMM/GMM^K	2.0	MFCCs +Δ+$\Delta\Delta$
SGMM^K	1.3	MFCCs +Δ+$\Delta\Delta$

Table 2: *Total number of parameters (N) for selected individual acoustic models implemented for both well- and under-resourced scenarios.*

are 13 dimensional MFCCs with per-speaker mean and variance normalization accompanied by first and second derivatives (MFCCs+Δ+$\Delta\Delta$). To evaluate the SGMM framework on more complex features, we decided to employ MLLT-MFCC features, similar to those of the benchmark system in the LVCSR task (employing HLDA transform over PLPs [8]). These features are represented by per-speaker normalized MFCCs spliced over 9 consecutive frames and projected by LDA (performing reduction to 40 dimensions). MLLT transform is then used over the LDA-reduced features.

To obtain ASR performance of the HMM/GMM on exactly the same training setup using Kaldi toolkit, first, the HMM/GMM system is implemented for the both well- and under-resourced ASR tasks. SGMM models are then trained with the same training setup, similar to the HMM/GMM. In addition, SGMM models are trained without and with speaker vectors, as described in Section 2.1. Although we do not apply any discriminative training procedure in any of our experiments, the Speaker Adaptive Training (SAT) is eventually applied in both SGMM and HMM/GMM systems (provided by feature-space adaptation using feature-space Maximum Likelihood Linear Regression (fMLLR), also known as constrained MLLR [12]).

4.1. Well-resourced LVCSR meeting system

As described in Section 3.1, HMM/GMM^K and SGMM^K systems, implemented using Kaldi, were first trained on well-resourced English meeting recordings available from AMI(DA) project[5]. The HMM/GMM^K uses around 4.5k tied-states and 100k Gaussian mixture components (in total). In case of the SGMM^K, an UBM is first initialized by clustering the diagonal Gaussian mixture components of the HMM/GMM^K system. The UBM is then trained with full-covariance matrices on the full training set. The final size of UBM is $I = 500$ Gaussians. Initialization of the SGMM^K model is done from the previously trained UBM with a sub-space dimension $S = 50$. The final SGMM^K contains 100k sub-states. In case of applying speaker vectors, size of the speaker subspace dimension is $T = 39$.

ASR results on well-resourced meeting data are reported in Table 1. WER of the HMM/GMM^H benchmark trained on exactly the same data is 41.7%. The second HMM/GMM^H benchmark trained on slightly less amount of data ($-1/5$) than in [8] and exploiting HLDA-PLP features gives 39.2%. The LM scaling factor is tuned in both systems for the best WER. There is no SAT applied in any of these systems. Kaldi-based HMM/GMM^K with simple MFCCs performs slightly worse than HTK-baseline. MLLT-MFCC features give about similar performance as HLDA-PLPs. Eventual SAT training reduces WER by 4.1%. The SGMM^K applying simple MFCCs (without using speaker vectors) gives WER equal to 38.2%. A reduction by about 2% absolute is obtained for the SGMM with speaker vectors. MLLT-MFCC features significantly improve WERs over simple MFCCs. Final WER after applying fMLLR feature transform is 34.2%. The overall number of parameters for selected acoustic models implemented on well-resourced meeting data is given in Table 2 (upper part).

4.2. Under-resourced small vocabulary ASR system on Afrikaans

Unlike LVCSR setup, in the second type of experiments, acoustic models were trained on relatively low amount (3 hours) of data. The Kaldi-based HMM/GMM^K contains around 1.8k tied-states and 25k Gaussian mixture components (in total). Similar to the LVCSR setup, an UBM initialized from the HMM/GMM^K, having $I = 400$ Gaussians, is used to initialize the SGMM^K. The final SGMM^K contains 7.5k sub-states and dimensions S and T are equal to 40.

Table 3 reports WERs obtained on under-resourced Afrikaans. Benchmark WER performance provided by KL-HMM^H and multilingual Tandem^H systems [1] is 37.2% and 38.9%, respectively. As mentioned in Section 3.2.1, these models benefit from additional out-of-language data. There is no SAT applied in any of these systems. Kaldi-based HMM/GMM^K trained only using in-domain data and applying similar kind of features (however MFCCs instead of PLPs) performs considerably worse (40.6%). MLLT-MFCC features reduce WER by 4.3% and SAT seems to perform very well (another 6.2% absolute reduction in WER). SGMM^K models were trained only using in-domain data. By using simple MFCCs, WER is 37.1%. By adding speaker vectors to the SGMM, WER is reduced by 3.1% absolute. MLLT-MFCCs and eventual fMLLR transform reduce WER down to 30.4%. The overall number of parameters for selected acoustic models developed on under-resourced Afrikaans is given in Table 2 (lower part).

[5] http://www.amidaproject.org

System	WER [%]	Description
Under-resourced small vocabulary ASR system on Afrikaans		
KL-HMM^H	37.2	Posterior features estimated on PLPs+Δ+$\Delta\Delta$
Tandem^H	38.9	Posterior features estimated on PLPs+Δ+$\Delta\Delta$
HMM/GMM^K	40.6	MFCCs +Δ+$\Delta\Delta$
HMM/GMM^K	36.3	MLLT-MFCCs
HMM/GMM^K	30.1	MLLT-MFCCs + SAT(fMLLR)
SGMM^K	37.1	MFCCs +Δ+$\Delta\Delta$
SGMM^K	33.9	MFCCs +Δ+$\Delta\Delta$ + speaker vectors
SGMM^K	30.8	MLLT-MFCCs + speaker vectors
SGMM^K	30.4	MLLT-MFCCs + speaker vectors + fMLLR

Table 3: *Experimental results of HMM/GMM and SGMM systems evaluated on under-resourced Afrikaans data. KL-HMM and Tandem benchmark systems, implemented using HTK ('H'), were developed for under-resourced scenarios and are boosted by additional data sources, as described in Section 3. All other systems, implemented using Kaldi ('K') are presented in Section 4.*

5. Conclusions and discussions

We have reported ASR performance of relatively new kind of acoustic models based on SGMM. So far achieved performance of SGMM models (in case of being purely trained using in-domain data) was reported only on standard (less ambitious) ASR datasets (Wall-Street Journal, Resource Management). Our SGMM experiments were carried out on two challenging but completely diverse corpora, i.e, well- and under-resourced datasets. For the sake of comparison, in addition to the SGMM, conventional HMM/GMM was trained in both cases using exactly the same amount of data. We also utilized the same kind of information during the training and evaluation (i.e., similar features, speaker adaptation, language model). Furthermore, for each scenario, interesting ASR performances reported by other authors were taken into account. In case of well-resourced data, HMM/GMMs trained on similar data were reported. In case of under-resourced corpora, state-of-the-art KL-HMM and multilingual Tandem systems were reported.

Experimental results clearly show that SGMM models (having usually less model parameters than other evaluated acoustic systems, as shown in Table 2) perform better for all the examined conditions (simple/complex features, speaker adaptation), and in both well and under-resourced scenarios.

6. Acknowledgment

The authors are grateful to the HLT group at Meraka, and especially to Dr. Febe de Wet, for providing us with the Afrikaans training and test sets as well as the Afrikaans dictionary.

7. References

[1] Imseng D., Bourlard H., Garner P., "Boosting under-resourced speech recognizers by exploiting out of language data - Case study on Afrikaans", to appear In Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages, 2012.

[2] Imseng D., Bourlard H., Garner P., "Using KL-divergence and multilingual information to improve ASR for under-resourced languages", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4869-4872, Kyoto, Japan, 2012.

[3] Povey D., et.al., "Subspace Gaussian Mixture Models for Speech Recognition", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4330-4333, Dallas, USA, 2010.

[4] Burget L., et.al., "Multilingual Acoustic Modeling For Speech Recognition Based On Subspace Gaussian Mixture Models", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4334-4337, Dallas, USA, 2010.

[5] Lin H., Deng L., Yu D., Gong Y., Acero A., Lee C., "A Study on Multilingual Acoustic Modeling for Large Vocabulary ASR", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4333-4336, Taipei, Taiwan, 2009.

[6] Povey D., et. al., "The Kaldi Speech Recognition Toolkit", In proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Hawaii, USA, 2011.

[7] Povey D., Karafiat M., Ghoshal A., Schwarz P., "A Symmetrization of the Subspace Gaussian Mixture Model", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4504-4507, Prague, Czech R., 2011.

[8] Grezl, F., Karafiat, M., Burget, L., "Investigation into bottle-neck features for meeting speech recognition", In Proceedings of Interspeech, pp. 2947-2950, Brighton, UK, 2009.

[9] Barnard E., Davel M. and Heerden C., "ASR Corpus design for resource-scarce languages", In Proceedings of Interspeech, pp. 2847-2850, Brighton, UK, 2009.

[10] Davel M. and Martirosian O., "Pronunciation dictionary development in resource-scarce environments", In Proceedings of Interspeech, pp. 2851-2854, Brighton, UK, 2009.

[11] Oostdijk N., "The spoken Dutch corpus – Overview and first evaluation", In Proceedings of the second International Conference on Language Resources and Evaluation, pp. 887-893, 2000.

[12] Gales M., "Maximum likelihood linear transformations for HMM-based speech recognition", In Computer Speech and language, pp. 75-98, April 1998.