INVESTIGATION OF ACOUSTIC MODELING TECHNIQUES FOR LVCSR SYSTEMS

X. Liu, M. J. F. Gales, K. C. Sim & K. Yu

Cambridge University Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {xl207,mjfg,kcs23,ky219}@eng.cam.ac.uk

ABSTRACT

This paper describes the use of several advanced acoustic modeling techniques for the 2004 CU-HTK large vocabulary speech recognition systems. These techniques include Gaussianization for speaker normalization, discriminative Cluster Adaptive Training (CAT), Subspace for Precision And Mean (SPAM) modeling of inverse covariances, and discriminative complexity control. Acoustic models featuring these techniques were integrated into a state-of-the-art 10 real-time multi-pass system with sophisticated adaptation for performance evaluation. Experimental results are presented on both broadcast news (BN) and conversational telephone speech (CTS) transcription tasks.

1. INTRODUCTION

For many years automatic transcription of broadcast news (BN) and conversational telephone speech (CTS) data have been the two main tasks for the research community of large vocabulary continuous speech recognition (LVCSR). Due to the difficulty of these tasks, a variety of modeling techniques have been developed to make these systems able to model more complex data and more robust to changes in acoustic environment. In this paper several advanced modeling techniques are investigated in the framework of a state-of-the-art multi-pass LVCSR system using sophisticated adaptation, large scale language models and Confusion Network (CN) based system combination. By implementing the approaches in this complex framework, it is possible to obtain a realistic comparison of how they may work in an evaluation style system. Techniques investigated include Gaussianization for speaker normalization, discriminative Cluster Adaptive Training (CAT), Subspace for Precision And Mean (SPAM) modeling of inverse covariances, and model complexity control.

The rest of the paper is organized as follows. Section 2 describes the four acoustic modeling techniques examined. Section 3 gives an overview of the basic features of the CU-HTK 10xRT system. Then experimental results of individual and combined systems on both BN and CTS transcription tasks are presented. Section 4 is the conclusion.

2. MODELING TECHNIQUES

This section describes the theory of Gaussianization, CAT, SPAM and discriminative complexity control. Some implementation issues are also discussed briefly for individual techniques.

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

2.1. Gaussianization

Cepstral mean and variance normalization is a simple speaker normalization scheme. The aim is to transform the distribution of a speaker's date to a zero mean and unit variance Gaussian. However such a simple linear normalization scheme may not be powerful enough to normalize highly non-homogeneous speech data, such as broadcast news. In this paper a non-linear speaker normalization scheme, *Gaussianization*, is investigated for both BN and CTS tasks. The basic idea is to ensure the Cumulative Density Function (CDF) of each speaker's data, which is represented by a Gaussian Mixture Model (GMM), matches that of a Gaussian distribution. Let o_j denote the jth dimension of a n dimensional acoustic feature vector o of speaker s. Then the Gaussianized feature on j the dimension is given by,

$$\breve{o}_j^{(s)} = \boldsymbol{\phi}^{-1}\left(\int_{-\infty}^{o_j^{(s)}} \sum_{m=1}^{M_{sj}} c_{sjm}\mathcal{N}\left(x;\, \mu^{(sjm)}, \sigma^{(sjm)}\right) \mathrm{d}x\right) \quad (1)$$

where $\boldsymbol{\phi}^{-1}(\cdot)$ denotes the standard Gaussian inverse CDF. The speaker GMM component mean, variance and prior is denoted by $\mu^{(sjm)}$, $\sigma^{(sjm)}$ and c_{sjm} respectively. For each speaker a total of n single dimension M_{sj} component GMMs were trained using Maximum Likelihood (ML) criterion. This scheme provides a compact and smooth representation of the target distribution, as a simplified version of iterative scheme proposed in [1]. It may be viewed as higher order version of cepstral mean and variance normalization. In this work Gaussianization was performed on top of HLDA projected cepstral features. The normalized features were then used in both training and testing. All GMMs used for Gaussianization had 4 components.

2.2. Cluster Adaptive Training

Multiple-cluster schemes, such as cluster adaptive training (CAT) or eigenvoices system, are popular approaches for rapid speaker and environment adaptation [3]. A multiple-cluster model is used as the canonical model in an adaptive training framework. A set of interpolation weights are used to transform this multiple-cluster model to a standard HMM set representative of an individual speaker or acoustic environment. Usually only multiple-cluster means are considered, thus adapted mean vector is represented as

$$\boldsymbol{\mu}^{(sm)} = \mathbf{M}^{(m)}\boldsymbol{\lambda}^{(s)} \quad (2)$$

where $\mathbf{M}^{(m)} = \left[\boldsymbol{\mu}_1^{(m)}, \ldots, \boldsymbol{\mu}_P^{(m)}\right]$ is the multiple-cluster mean matrix, and $\boldsymbol{\lambda}^{(s)}$ is the interpolation weight vector.

In *Proc. ICASSP 2005* Volume I, pp. 849 – 852, March 2005, Philadelphia.

Maximum likelihood estimation for the multiple-cluster model and interpolation weights are investigated in [3]. Initializations of CAT is also detailed discussed in the paper, which allows CAT to be used in LVCSR systems. However, to get state-of-the-art performance, discriminative training, particularly minimum phone error (MPE) training is required. This has been studied for multiple cluster systems in [5]. Though both model parameters and interpolation weights can be discriminatively updated, a simplified version of discriminative adaptive training is commonly used, in which ML-estimated weights are fixed in later discriminative training stage.

In the CU-HTK 10xRT system, a CAT system employs similar adaptation procedures to a SAT system. CAT weights are iteratively estimated using ML criterion based on supervision from the previous lattice generation stage. Then given using these transforms, standard MLLR transforms are estimated in a cascade fashion for lattice rescoring.

2.3. Precision Matrix Modeling

Standard GMMs for speech recognition use diagonal covariance matrices. Structured precision matrix approximations have been found to yield improved performance using both ML and MPE training [6, 7]. They yield a compact representation and efficient likelihood calculation. Examples of this form of model are the Semi-tied Covariances (STC), Extended Maximum Likelihood Linear Transform (EMLLT) and Subspace for Precision And Mean (SPAM) systems. The precision matrix (inverse covariance), $\boldsymbol{P}_m$, of a Gaussian component m, can be expressed in a general form of basis superposition:

$$\boldsymbol{P}_m \quad = \quad \sum_{i=1}^{n} \lambda_{ii}^{(m)} \boldsymbol{S}_i \tag{3}$$

where $\boldsymbol{S}_i$ is the ith basis matrix and $\lambda_{ii}^{(m)}$ is the corresponding basis coefficient. This form of precision matrix model has been found to yield the best performance [7].

This paper considers MPE discriminatively trained SPAM models. Two variants of SPAM models were trained. The first model was trained within the 39-dimensional HLDA feature space. The second form of model was built with an adaptively trained feature-space. Here constrained MLLR was used to generate a standard ML Speaker Adaptively Trained (SAT) system. Then within the adaptively trained feature-space the precision matrix models were built. This is the SAT-SPAM system

One issue with using structured precision matrix models is that if the adaptation transforms are directly estimated, this can be computationally expensive, or require numerical optimization techniques. In this work the efficient adaptation schemes described in [10] were used. A diagonal precision matrix approximation was used for MLLR mean adaptation. For constrained MLLR (CMLLR), a standard diagonal covariance matrix system was used to estimate the CMLLR transforms.

2.4. Complexity Control

There are a wide range of possible models that can be used for LVCSR. One issue is that it is very expensive to build, and compare, each possible system. To overcome this problem automatic model complexity control schemes may be used. Most existing complexity control schemes make an assumption that increasing the likelihood on held-out data can decrease the word error rate (WER). However this correlation has been found quite weak for current speech recognition systems. It would be preferable to use a criterion more closely related to WER. One possible method is to marginalize a discriminative criterion. However, due to sensitivity to outliers, discriminative training criteria, such as Maximum Mutual Information (MMI), can not be directly integrated for complexity control.

To overcome this problem the marginalization of a discriminative growth function has been proposed [11]. Let λ denotes the model parameters. For a family of discriminative criteria that can be expressed as a ratio between two polynomials with positive coefficients (including MMI and MPE), $\mathcal{F}(\lambda) = \mathcal{F}_{\mathrm{num}}(\lambda)/\mathcal{F}_{\mathrm{den}}(\lambda)$, a generic form of the associated growth function is given below.

$$\mathcal{G}(\lambda) \quad = \quad \mathcal{F}_{\mathrm{den}}(\lambda) \left[\mathcal{F}(\lambda) - \mathcal{F}(\tilde{\lambda}) + C\mathcal{F}_{\mathrm{sm}}(\lambda, \tilde{\lambda}) \right] \tag{4}$$

where $\tilde{\lambda}$ is the *current* parameter estimate. The first two terms in the bracket retain the criterion's curvature in the parametric space. A third smoothing criterion or statistics, $\mathcal{F}_{\mathrm{sm}}(\lambda, \tilde{\lambda})$, scaled by a constant $C > 0$, acts to remove the sensitivity to outliers by by penalizing highly unlikely word sequences. The exact form of the smoothing term depends on the underlying discriminative criterion being considered. Using a generalized EM approach, a strict lower bound of the growth function can be derived. This has a more tractable form for marginalization, with the dependence on the hidden variables removed. A second order Laplace's approximation can be used for the growth function integration.

In this paper complexity controlled acoustic models were built using this marginalized growth function. Two forms of complexity were varied. In contrast to the standard global 39-dimension HLDA projection, the systems were built with multiple HLDA transforms, in this case 65, with number of retained dimensions varied. In addition the number of components per state were varied. Both forms were determined using a marginalized MPE criterion. The BN complexity control system had 16.5 components per state and 46.3 dimensions per HLDA transform on average. The corresponding CTS system had 29.9 Gaussians per state and 42.6 dimensions per HLDA projection.

3. EXPERIMENTS AND RESULTS

3.1. Basic Features of CU-HTK Systems

The CU-HTK 10xRT multi-pass system uses sophisticated adaptation and CN based system combination. The overall system structure consists of two main stages: the initial lattice generation stage and the rescoring stage using multiple model sets. The confusion network outputs from different rescoring passes were finally combined. This is shown in figure 1. More details of the overall system architecture can be found in [2].

For both systems the audio data is parameterized using 13 PLP features augmented with their first, second and third order derivatives. A 52 dimensional acoustic feature was projected down to 39 dimension using a global HLDA transform. All acoustic models were built using discriminative training based on the minimum phone error (MPE) criterion [4]. For the CTS systems only, Vocal Tract Length Normalization (VTLN) was used in training and test and Cepstral mean and variance normalization was also applied. Continuous density, mixture of Gaussians, cross-word triphone HMM systems were used. Bandwidth-specific acoustic models

were also built for BN data. In addition Gender-specific BN models were derived from the gender-independent models. All CTS acoustic models were gender independent. The two baseline models sets used in the lattice rescoring stage are a SAT model employing constrained MLLR and an HMM set trained using a Single Pronunciation (SPron) dictionary. These model sets were adapted using lattice based MLLR in addition to standard adaptation only based on the 1-best hypothesis.

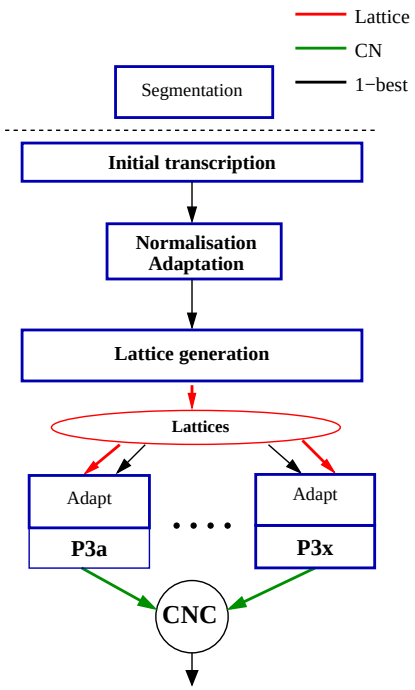

Fig. 1. CU-HTK 10xRT System

For both BN and CTS tasks a word-based 4-gram language model was trained on the acoustic transcriptions and additional Broadcast News data. The word-based 4-gram was then interpolated with a class-based trigram trained only on the associated acoustic transcriptions. The BN and CTS recognition dictionaries contain approximately 59k and 58k words respectively. Each word had about 1.1 pronunciations on average for both tasks.

3.2. CTS Experiments

The CTS data set used for training, `fsh2004sub`, consists of 400 hours of Fisher conversations released by the LDC, with a balanced gender and line condition [8]. Quick transcriptions are provided by BBN, LDC and another commercial transcription service. Two CTS test sets were used for systems evaluation. A 6 hour DARPA RT-03 evaluation set, `eval03`, contains 72 conversations from the LDC Fisher collection, fsh, and Switchboard II phase 5, s25. Another DARPA development set `dev04` was also used, which includes 72 LDC released Fisher conversations. All CTS models have approximately 6k physical states after decision tree based tying. The number of components per state is 28 on average level.

Table 1 shows the baseline performance of the 10 time real-time CTS system. The 2-way combination between the SAT and SPron systems was the standard configuration used in the CUED CTS evaluation system. Significant error rate reduction over individual branches was achieved after system combination. The final error rates were 20.5% on `eval03` and 16.9% on `dev04`.

System		eval03			dev04
		s25	fsh	Avg	
P2-cn	HLDA	26.6	18.4	22.6	18.7
P3a-cn	SAT	24.5	17.1	20.9	17.3
P3c-cn	SPron	24.7	17.6	21.3	17.6
P3a+P3c		23.9	16.8	20.5	16.9

Table 1. CTS 10xRT system baseline performance

Table 2 shows the performances of various systems featuring techniques described in section 2. The global HLDA system used for lattice generation was also re-adapted as a rescoring branch. The Gaussianization (GAUSS) and complexity controlled system (CTRL) systems gave marginal improvement. The SPAM system gave 0.8% absolute improvement on `eval03` over the HLDA baseline. An absolute word error reduction of 0.3% was also obtained on `dev04` against the P3b branch. Among all the adaptively trained systems, the SAT+SPAM outperformed all the other systems on both test sets. An absolute WER reduction of 0.4%∼0.5% were obtained on both sets over the SAT branch. The performances of SAT and CAT systems are very close.

System		eval03			dev04
		s25	fsh	Avg	
P3b-cn	HLDA	24.8	17.7	21.4	17.5
P3d-cn	GAUSS	24.8	17.5	21.3	17.3
P3e-cn	CAT	24.9	17.2	21.2	17.5
P3g-cn	SPAM	24.1	16.9	20.6	17.2
P3h-cn	SAT+SPAM	23.9	16.9	20.5	16.8
P3i-cn	CTRL	24.5	17.5	21.1	17.6
P3c+P3h		23.6	16.4	20.1	16.6
P3c+P3d+P3h		23.6	16.4	20.1	16.5
P3c+P3h+P3i		23.3	16.3	19.9	16.6

Table 2. Extended CTS 10xRT system performance

The GAUSS, SAT+SPAM and CTRL systems were then used for combination with the SAT and SPron systems. Replacing the SAT system with the SAT+SPAM branch reduced the error rate by 0.4% on `eval03` and 0.3% on `dev04`. Adding the GAUSS system in a 3-way combination with the SPron and SAT+SPAM branches gave further marginal gain on `dev04`. Similarly the error rate on `eval03` was reduced by 0.2% using a 3-way combination by further including the CTRL system.

3.3. BN Experiments

The BN system was trained on 370 hours of training data. This consists of two parts [9], 140 hours of accurately transcribed broadcast news acoustic training data released by the LDC in 1996 and 1997 and 230 hours of data selected from the TDT4 audio corpora with close-captions based quick transcriptions. All BN models have approximately 7k physical states after decision tree based tying. The number of components per state is 16 on average level.

Three BN test sets were used, each of them contains six 30 minutes broadcast news shows. The first set, `eval03`, was the DARPA RT-03 evaluation data set. It contains shows which were broadcast during February 2001. Two additional DARPA internal development sets, `dev04` and `dev04f` were also used. They contain shows of January 2001 and November 2003 respectively.

System		eval03	dev04	dev04f
P2-cn	HLDA	10.8	13.4	20.1
P3a-cn	SAT	10.3	12.9	18.7
P3c-cn	SPron	10.2	13.0	19.0
P2+P3a+P3c		10.1	12.6	18.6

Table 3. BN 10xRT system baseline performance

Table 3 shows the performance of the baseline BN 10xRT system. In contrast to the CTS system, a 3-way combination between the P2, P3a (SAT) and P3c (SPron) branches was the standard configuration used in CUED BN evaluation system. The final numbers for each of the tasks was 10.1%, 12.6% and 18.6%, with gains of 0.1% to 0.4% being obtained from system combination.

System		eval03	dev04	dev04f
P3b-cn	HLDA	10.5	13.1	19.5
P3d-cn	GAUSS	10.4	12.8	19.1
P3e-cn	CAT	10.4	12.8	19.1
P3g-cn	SPAM	10.2	12.7	18.8
P3h-cn	SAT+SPAM	10.1	12.5	18.5
P3i-cn	CTRL	10.5	12.8	19.3
P2+P3c+P3f		10.0	12.5	18.6
P2+P3c+P3h		10.0	12.4	18.4
P2+P3a+P3c+P3h		10.0	12.4	18.4

Table 4. Extended BN 10xRT system performance

Table 4 shows the performances of various BN systems. The Gaussianization system outperformed the HLDA system on all three sets. 0.3%$\sim$0.4% error rate reduction is obtained on `dev04` and `dev04f`. The SPAM system was the best non-adaptively trained system, by an absolute WER reduction of 0.3%$\sim$0.7% against the P3b system. Performances of the two SPAM systems are close. The CAT system consistently outperformed the HLDA baseline system on all sets, while marginal gain was found over the SAT system. The gain from the CTRL system over the HLDA baseline was marginal similar to the CTS experiments in table 2. The SAT+SPAM system was then selected for combination. Using the SAT+SPAM branch reduced the error rate by 0.1% on `eval03` and 0.2% on both `dev04` and `dev04f`. Unfortunately further including the SAT branch in a 4-way combination with the p2, SPron and SAT+SPAM systems gave the same performance.

4. CONCLUSION

In this paper several advanced acoustic modeling techniques, Gaussianization, CAT, SPAM and complexity control were investigated for LVCSR training. Performances of individual and combined systems were compared in the framework of a state-of-the-art 10 time real time system for both BN and CTS data. Experimental results show that these techniques are useful for further improving performance of current LVCSR systems.

5. REFERENCES

[1] G. Saon, A. Dharanipragada, & D. Povey (2004). Feature Space Gaussianization, *Proc. ICASSP'04*, Montreal.

[2] G. Evermann, & P. C. Woodland (2003). Design of Fast LVCSR Systems. *Proc. ASRU'03* , St. Thomas, U.S. Virgin Islands.

[3] M. J. F. Gales (2000). Cluster Adaptive Training of Hidden Markov Models, *IEEE Transactions on Speech and Audio Processing*, pp. 417–428, Vol. 8, 2000.

[4] D. Povey & P. C. Woodland (2002). Minimum Phone Error and I-smoothing for Improved Discriminative Training, *Proc. ICASSP'02*, Florida, USA.

[5] K. Yu & M. J. F. Gales (2004). *Discriminative cluster adaptive training*, Cambridge University Engineering Department Technical Report, CUED/F-INFENG/TR-486, August 2004.

[6] S. Axelrod, V. Goel, B. Kingsbury, K. Visweswariah, & R. A. Gopinath (2003), Large Vocabulary Conversational Speech Recognition with a Subspace Constraint on Inverse Covariance Matrices, *Proc. Eurospeech'03*.

[7] K. C. Sim & M. J. F. Gales, *Precision matrix modelling for large vocabulary continuous speech recognition*, Cambridge University Engineering Department Technical Report, CUED/F-INFENG/TR-485, July 2004.

[8] G. Evermann, H. Y. Chan, M. J. F. Gales, B. Jia, D. Mrva, P.C. Woodland & K. Yu (2004). Training LVCSR Systems on Thousands of Hours of Data, Summited to *ICASSP'05*.

[9] D. Y. Kim, M. J. F. Gales, G. Evermann, H. Y. Chan, K. C. Sim, D. Mrva & P. C. Woodland (2004). Development of the CU-HTK 2004 Broadcast News Transcription System, Summited to *ICASSP'05*.

[10] K. C. Sim & M. J. F. Gales (2004). Adaptation of Precision Matrix Models on Large Vocabulary Speech Recognition, Summited to *ICASSP'05*.

[11] X. Liu & M. J. F. Gales (2004). *Automatic Model Complexity Control Using Marginalized Discriminative Growth Functions*, Cambridge University Engineering Department Technical Report, CUED/F-INFENG/TR-490, August 2004.