PERFORMANCE COMPARISONS OF ALL-PASS TRANSFORM ADAPTATION WITH MAXIMUM LIKELIHOOD LINEAR REGRESSION

John McDonough and Alex Waibel

Interactive Systems Laboratories
Institut für Logik, Komplexität, und Deduktionssysteme
Universität Karlsruhe
Am Fasanengarten 5, 76128 Karlsruhe, Germany
jmcd@ira.uka.de
http://isl.ira.uka.de/~jmcd

ABSTRACT

All-pass transform (APT) adaptation transforms the cepstral means of a hidden Markov model so as to mimic the effect of warping the short-time frequency axis of a segment of speech, much like vocal tract length normalization (VTLN). APT adaptation can be implemented as a linear transformation in the cepstral domain, however, much like the better known maximum likelihood linear regression (MLLR). Recent work demonstrated the superior performance of APT adaptation to MLLR for a large vocabulary conversational speech recognition task. This work presents similar comparisons on the *Switchboard Corpus*. We found that *without* VTLN, the best MLLR and APT systems achieved word error rates (WERs) of 43.0% and 40.2% respectively. Similarly, *with* VTLN the respective error rates were 40.3%, and 39.2%, so that APT adaptation is significantly better in both cases. We also undertook a set of experiments to determine whether APT adaptation can be combined with a linear semi-tied covariance (STC) transform. With a single APT per speaker, the application of STC reduced the WER from 42.9% to 39.4%.

1. INTRODUCTION

All-pass transform (APT) adaptation transforms the cepstral means of a hidden Markov model (HMM) so as to mimic the effect of warping the short-time frequency axis of a segment of speech [8], much like vocal tract length normalization (VTLN) [2]. APT adaptation can be implemented as a linear transformation in the cepstral domain, however, much like the better known maximum likelihood linear regression (MLLR) [6].

Speaker-adapted training (SAT) is an algorithm for performing maximum likelihood estimation of the parameters of a continuous density HMM when speaker adaptation is applied during both training and test [1]. SAT can be used with any speaker adaptation scheme employing a linear transformation of cepstral means, including both MLLR and APT adaptation. In a typical implementation of speaker adaptation, the Gaussian components of an HMM are partitioned into a number of mutually exclusive sets or *regression classes*; several straightforward modifications of the basic SAT algorithm have been proposed to update the assignment of Gaussian components to classes using a maximum likelihood (ML) criterion. Single-pass adapted training (SPAT) [8], is a variation of SAT tailored specifically for use with APT-based adaptation. SPAT makes extensive use of an HMM with one Gaussian component per state cluster to estimate speaker-dependent APT parameters; these parameters are then transferred to the final multiple-mixture HMM in a computationally-efficient manner. Incremental training (IT) [9] is a refinement of SPAT that gradually adds adaptation parameters for improved recognition performance.

Using large vocabulary conversational speech recognition (LVCSR) systems trained and tested on speech material from the *Switchboard Corpus*, we compare the reduction in WER provided by APT adaptation and IT with that achieved with MLLR and standard SAT. These experiments complement those described in [7], in which APT adaptation proved to provide superior performance to MLLR on another LVCSR task. We also investigate the combination of APT adaptation with the linear feature transformation provided by semi-tied covariance (STC) estimation [4].

The balance of this work is organized as follows. In Setion 2 we briefly review the definition of the APT. In Section 3 we present the results of several series of speech recognition experiments. These include performance comparisons of APT adaptation with MLLR, both without and with VTLN in Sections 3.1 and 3.2, respectively. We also describe in Section 3.3 a series of experiments undertaken to determine whether APT adaptation can be successfully combined with the linear transformation of cepstral features that follows with the use of STC matrices. Section 4 presents our plans for future work.

2. REVIEW OF ALL-PASS TRANSFORMS

Here we briefly review the properties of the *all-pass transform* (APT), which is used to formulate a basis for speaker adaptation. An APT is defined as

$$Q(z) = z \exp F(z) \tag{1}$$

where

$$F(z) = \sum_{k=1}^{M} \alpha_k F_k(z) \text{ for } \alpha_1, \ldots, \alpha_M \in R, \tag{2}$$

$$F_k(z) = \frac{\pi}{2}\left(z^k - z^{-k}\right) \tag{3}$$

This work was sponsored by the FAME project.

and M is the number of free parameters $\{\alpha_k\}$ in the transform. It can be readily verified that Q as defined (1) is an *all-pass* function [10, §5] in the sense that it preserves the unit circle. Moreover, as z traverses the unit circle once, $Q(z)$ also winds exactly *once* about the origin, which is necessary to ensure that spectral content is not doubled or tripled [8, §3.5].

In order to calculate the coefficients of a transformed cepstral sequence, it is first necessary to calculate the coefficients q in the Laurent series expansion of Q; this can be done as follows: For F as in (2) set

$$G(z) = \exp F(z) \tag{4}$$

and let g denote the coefficients of the Laurent series expansion of G valid in an annular region including the unit circle. The sequence f of coefficients in the series expansion of F are available by inspection from (2) and (3). It can then be shown [8, §3.5] that

$$g[n] = \sum_{m=0}^{\infty} \frac{1}{m!} f^{(m)}[n]$$

Moreover, from the Cauchy product it follows $f^{(m)} = f * f^{(m-1)}$ for $m = 2, 3, \ldots$. Equation (4) implies that $Q(z) = z\,G(z)$, so the desired coefficients are given by $q[n] = g[n-1]$ for all $n = 0, \pm 1, \pm 2, \ldots$.

Let us define the *transformation matrix* $A = \{a_{nm}\}$ whose individual components are given by [8, §3.3]

$$a_{nm} = \begin{cases} q^{(m)}[0], & \text{for } n = 0, m \geq 0 \\ 0, & \text{for } n > 0, m = 0 \\ \left(q^{(m)}[n] + q^{(m)}[-n]\right), & \text{for } n, m > 0 \end{cases} \tag{5}$$

where the sequences $q^{(m)}$ are defined through the recurrence relation $q^{(m)} = q * q^{(m-1)}$. Assuming the basic APT is augmented by an offset vector b, the adapted mean $\hat{\mu}_k$ of the k^{th} Gaussian component of an HMM can be obtained from the speaker-independent (SI) mean μ_k according to

$$\hat{\mu}_k = A\mu_k + b \tag{6}$$

Note that A as defined in (5) has an infinite number of columns and must be truncated. This implies that μ_k can be *extended* to any desired length in the normal course of speaker-adapted training [7].

3. SPEECH RECOGNITION EXPERIMENTS

The speech experiments described below were conducted with the Janus Recognition Toolkit (JRTk), which is developed and maintained jointly at the University of Karlsruhe, in Karlsruhe, Germany and at the Carnegie Mellon University in Pittsburgh, Pennsylvania, USA.

For the experiments reported below, HMM training was conducted on 170 hours of speech material extracted from the Switchboard Corpus, which was donated by 580 speakers in 2,761 conversation sides. The test set used was that defined for the EARS Project 2002 Dry Run, which consisted of 12,863 words of telephone and cell phone speech in a total of 12 conversations. The baseline model contained 6,182 codebooks, each with varying numbers of Gaussian components, for a total of 168,398 Gaussians. Mean-subtracted length-13 cepstral features, along with delta and delta-delta features, were used for training and test, for a final feature length of 39. Conventional FFT-domain vocal tract length normalization (VTLN) based on fixed warp factors estimated with a previous system was used for those cases requiring it.

No. Regression Classes	% Word Error Rate
Unadapted Baseline	55.6 (53.9)
1	45.2 (43.3)
2	45.4 (43.0)
3	45.3 (43.1)
6	46.8 (44.2)

Table 1. Results of unsupervised adaptation with full-matrix MLLR without VTLN.

In reporting the word error rate (WER) results of the experiments described below, our custom will be to quote the single-best recognition WER, as well as the best WER obtained by optimizing the language model weight, word insertion penalty and silence weight based on the correct transcription. The latter will appear in parentheses. Given that this optimized WER is typically more indicative of true system performance, we will use it as the basis of the conclusions drawn in the course of this work.

3.1. Experiments without VTLN

In the initial set of experiments, we wished to compare the effectiveness of APT adaptation against that of MLLR in reducing word error rate for the Switchboard task when VTLN was used for *neither* training nor test. For these experiments, we first performed K-means followed by conventional label training to obtain the baseline model. This baseline model without speaker adaptation achieved a WER of 55.6% (53.9%). During unadapted decoding, word lattices were written, which were then used for three iterations of unsupervised adaptation with the multiple-mixture models, as described in [11]. To speed-up convergence in estimating APT parameters, the best parameters from the previous model of lower complexity were used as an initial point. A scale factor of 0.040 was applied to the acoustic log-likelihoods of all models prior to their combination with the unweighted language model log-likelihoods. This measure was intended to prevent the single-best Viterbi path, which is quite likely to contain errors, from dominating the statistics required for speaker adaptation.

The conventionally-trained model was tested with MLLR adaptation and achieved a word error rate of 52.5% (50.1%). Beginning from the conventionally-trained model, speaker-adapted training was performed with MLLR adaptation and varying numbers of regression classes. The systems based on APT adaptation were incrementally-trained as described in [9] with a frame count threshold of 50.0. Tables 1 and 2 provide the WER results for MLLR and APT adaptation respectively. In Table 2, APT-M indicates the use of an APT as in (2) with M free parameters. For the first two APT results reported in Table 2, both the original μ_k and transformed $\hat{\mu}_k$ means in (6) had a length of 39. For the latter results μ_k was extended to a length of 78 during SAT. From these results, it is clear that the MLLR-adapted systems obtained with the SAT procedure outperform the adapted conventionally-trained model, and that both MLLR and APT adaptation provide large reductions in WER. The best MLLR system had 2 RCs and achieved a WER of 45.4% (43.0%), which was significantly worse than the best APT system, which had 12 RCs and achieved a WER of 41.9% (40.2%). Without VTLN, the superiority of APT adaptation to MLLR is indisputable.

Train/Test Condition	% Word Error Rate
APT-1, 1 RC	49.1 (47.8)
APT-9, 1 RC	45.7 (44.8)
APT-9, 1 RC, mean len. = 78	43.7 (42.6)
APT-9, 4 RCs, mean len. = 78	43.0 (41.6)
APT-9, 8 RCs, mean len. = 78	42.5 (41.1)
APT-9, 12 RCs, mean len. = 78	41.9 (40.2)

Table 2. Results of unsupervised APT adaptation without VTLN.

No. Regression Classes	% Word Error Rate
Unadapted Baseline	50.9 (48.8)
1	42.6 (40.3)
2	42.7 (40.5)
3	43.2 (40.5)
6	44.4 (41.2)

Table 3. Results of unsupervised adaptation with MLLR and VTLN.

3.2. Experiments with VTLN

As is well-known, VTLN has proven useful in reducing word error rate for many LVCSR tasks, including Switchboard. Hence, a second set of experiments was undertaken to determine relative effectiveness of MLLR and APT adaptation when combined with VTLN. For these experiments, fixed VTLN warp factors, which had been previously estimated with the best current system under a maximum likelihood criterion, were used for both training and test. New cepstral mean vectors were then estimated based on the best speaker-dependent warp factors. Employing the warp factors and new mean vectors, conventional, MLLR speaker-adapted, and APT incremental training were conducted exactly as before. The test conditions were also the same as before. The unadapted conventionally-trained system *with* VTLN was used to write word lattices, which were subsequently used for unsupervised speaker adaptation. The conventionally-trained system achieved a WER of 50.9% (48.8%) unadapted and 48.6% (45.8%) with MLLR adaptation.

The systems used for the APT adaptation experiments were once more trained trained with the IT procedure described in [9], with a frame count threshold of 50.0 during training and test. The results of the MLLR and APT adaptation experiments are summarized in Tables 3 and 4, respectively. The best MLLR system had a single RC and achieved a WER of 42.6% (40.3%), while the best APT system had 16 RCs and achieved a WER of 41.5% (39.9%). Hence, APT adaptation was still marginally better than MLLR. Note that the performance of the MLLR systems improved significantly with the use of VTLN. For APT adaption, the use of VTLN together with the simple APT-1 transformation provided a reduction in WER with respect to the no VTLN case; compare 49.1% (47.8%) for APT-1 without VTLN to 47.2% (45.6%) with VTLN. But the WER reduction afforded by VTLN quickly vanished with the use of the more complicated APT-9 transformation and the addition of more regresion classes.

From Table 4 it is clear that APT adaptation fails to provide further reductions in WER with the addition of more than 16 RCs. Based on this failure, we suspected that perhaps the frame count threshold of 50.0 was too low. Hence, we repeated the entire IT procedure with a threshold of 400.0, and tested the new models with a threshold of both 150.0 and 400.0. The results of these tests are shown in Table 5, in which the first column is repeated from Table 4. Clearly, the use of a threshold of 150.0 had a beneficial effect, as the best system with this threshold had 44 RCs and achieved a WER of 41.2% (39.4%), a significant improvement over the best MLLR system. Moreover, we observed that the addition of more RCs only ceased to provide further reductions in WER when the amount of *test* data was no longer sufficient to perform unsupervised parameter estimation on all or most leaf nodes of the regression tree, and backing off became universal. Hence we surmise that further reductions in WER would be possible, were the test conversations of longer duration, thereby providing more adaptation material.

Train/Test Condition	% Word Error Rate
Unadapted Baseline	50.9 (48.8)
APT-1, 1 RC	47.2 (45.6)
APT-9, 1 RC	44.8 (43.8)
APT-9, 1 RC, mean len. = 78	43.9 (42.9)
APT-9, 4 RCs, mean len. = 78	42.7 (41.2)
APT-9, 8 RCs, mean len. = 78	42.2 (40.8)
APT-9, 12 RCs, mean len. = 78	41.6 (40.6)
APT-9, 16 RCs, mean len. = 78	41.5 (39.9)
APT-9, 24 RCs, mean len. = 78	41.5 (40.0)
APT-9, 32 RCs, mean len. = 78	41.2 (40.0)

Table 4. Results of unsupervised APT adaptation with VTLN.

No. Reg. Classes	Adaptation Threshold		
	50.0	150.0	400.0
8	42.2 (40.8)	41.8 (40.7)	N/A
12	41.6 (40.6)	41.3 (40.3)	41.2 (40.3)
16	41.5 (39.9)	41.1 (40.2)	41.1 (40.3)
24	41.5 (40.0)	41.5 (39.9)	41.5 (39.9)
32	41.2 (40.0)	41.1 (39.5)	41.4 (39.7)
44	N/A	41.2 (39.4)	41.7 (40.1)
56	N/A	42.2 (40.8)	41.5 (39.8)

Table 5. Results of unsupervised 9-parameter APT adaptation with VTLN.

In a final experiment, we used the best 44-regression class system from above as a starting point, then increased the number of free parameters in each APT transform from 9 to 17. The resulting system achieved a WER of 41.0% (39.2%).

3.3. Experiments with STC

In the recent past, it has become increasingly popular to apply one or more linear transformation to the raw cepstral features *prior* to their use in speech recognition. Among the linear transformations that have proven useful for this application are traditional *linear discriminant analysis* (LDA) as discussed in [3, §10], a variant of traditional LDA known as *heteroscedastic linear discriminant analysis* (HLDA) proposed by Kumar and Andreou [5], and *semi-tied covariance* (STC) transformations as proprosed by Gales [4]. Because the formulation of APT adaptation exploits the characteristics of cepstral sequences, APT adaptation must always be applied prior to any other transformation. But given the

Adaptation Condition	STC Condition			
	No STC	Case A	Case B	Case C
APT-1, 1 RC	47.2 (45.6)	46.2 (45.5)	45.9 (45.3)	44.3 (43.7)
APT-9, 1 RC	44.8 (43.8)	43.7 (43.1)	43.4 (42.8)	41.8 (41.2)
APT-9, 1 RC, mean len. = 78	43.9 (42.9)	42.1 (41.6)	41.6 (41.1)	39.9 (39.4)
APT-9, 4 RCs, mean len. = 78	42.7 (41.2)	41.9 (41.0)	41.0 (40.2)	39.0 (38.5)

Table 6. Results of unsupervised 9-parameter APT adaptation with VTLN and STC.

sizeable reductions in WER provided by the transformations mentioned above, we deemed it worthwhile to combine the benefits of APT adaptation and linear feature transformation. Hence, we undertook a final set of experiments to determine whether APT adaptation could be combined with with STC, a representative linear transformation. We considered several distinct cases:

- **Case A:** A global STC transformation was estimated during SAT, but the APT parameters for training and test were retained from the non-STC experiments.
- **Case B:** Both the global STC transformation and speaker-dependent APT parameters were estimated during SAT. The global STC transformation was held fixed for test, while new APT parameters for each test set speaker were obtained with unsupervised lattice adaptation.
- **Case C:** APT parameters as well as *speaker-dependent* STC transformations were estimated for each speaker during SAT; unsupervised lattice adaptation was used to estimate both APT and STC transformation parameters for each test set speaker.

The results of these experiments are shown in Table 6.

From the results above, we observe that the use of STC feature normalization yields a word error rate reduction in all cases. Comparing Case A with Case B, we also see that the joint estimation of STC and APT parameters is better than holding the APT parameters estimated during non-STC training fixed and simply estimating the STC transformation based on them. As is clear upon examination of Case C, the very best use of STC comes from the estimation a unique transformation matrix for each speaker. A tentative explanation for this fact can be stated as follows: APT adaptation transforms the SI means of an HMM so as to better match the cepstral features of a given speaker. In so doing, however, the means are transformed into a space that may not have the desired diagonal covariance structure assumed by the HMM. Hence, a second transformation is necessary to transform both features and adapted means back into this "diagonal covariance" space. Because the cepstral features and APT parameters are unique for each speaker, this diagonalizing transformation must also be individually estimated for each speaker.

From the final row in Table 6, we observe that the combined APT/STC training scheme is apparently not optimal for the use of multiple APT transformations per speaker. As a corollary of the argument above, it may in fact be necessary to estimate multiple STC transformations per speaker when multiple APT transformations are used.

4. FUTURE WORK

Future work will concentrate on further refinements of the incremental training procedure described in [9]. Specifically, we must determine whether the regression class splitting procedure used in this work is optimal. It would also be of interest to compare the performance of APT adaptation to MLLR and other forms of speaker adaptation on a task where dozens or hundreds of minutes of unsupervised enrollment data is available for each speaker, as in the recognition of lectures, speeches, or meetings, for example. In this work, we made an initial step in combining APT adaptation with STC [4], one of the several currently-popular linear feature transformations. Further work is necessary to determine if APT adaptation can also be used with LDA [3, §10] or HLDA [5].

5. REFERENCES

[1] A. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *Proc. ICSLP*, 1996.

[2] Andreas Andreou, Theresa Kamm, and Jordan Cohen. Experiments in vocal tract normalization. In *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[3] Keinosuke Fukunaga. *Statistical Pattern Recognition*. Academic Press, San Diego, second edition, 1990.

[4] M. J. F. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Transaction on Speech and Audio Processing*, 7:272–281, 1999.

[5] N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, 26:283–297, 1998.

[6] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, pages 171–185, 1995.

[7] John McDonough, Thomas Schaaf, and Alex Waibel. Speaker adaptation with all-pass transforms. *Speech Communication, Special Issue on Adaptation Methods for Speech Recognition*, 42:75–91, January 2004.

[8] John W. McDonough. *Speaker Compensation with All-Pass Transforms*. PhD thesis, The Johns Hopkins University, Baltimore, MD, 2000.

[9] John W. McDonough. Performance comparisons of all-pass transform adaptation with maximum likelihood linear regression. Technical Report 102, Interactive Systems Laboratories, University of Karslruhe, 2003.

[10] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

[11] L. Uebel and P. Woodland. Improvements in linear transform based speaker adaptation. In *Proc. ICASSP*, 2001.