ROBUST LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION USING POLYNOMIAL SEGMENT MODEL WITH UNSUPERVISED ADAPTATION

Man-Hung SIU, and Siu-Kei AU YEUNG

Department of Electrical and Electronic Engineering
Hong Kong University of Science and Technology, Hong Kong
eemsiu@ust.hk, eejeffay@ust.hk

ABSTRACT

Robustness has been an important issue for applying speech technologies to real applications. While the Polynomial Segment Models (PSMs) have been shown to outperform HMM under the clean environment, the segmental likelihood evaluation may make the PSM distributions sharper and may adversely affect their performance in mis-matched conditions. In this paper, we explore the robustness properties of the PSM under noisy and channel mis-match conditions. In addition, unsupervised adaptation techniques have been shown to work well for environmental adaptation even with small amount of adaptation data. Thus, it is interesting to compare the PSMs' and the HMMs' performances after applying two types of unsupervised adaptation: the Maximum Likelihood Linear Regression (MLLR) and the Reference Speaker Weighting (RSW). Experiments were performed on the Aurora 4 corpus under both clean and multi-conditional training. Our results show that even under noisy and mis-match conditions, the PSMs performed well compared to the HMMs both before and after environmental adaptation. Using the best lattice, the RSW adapted PSM gave word error rates of 26.5% and 21.3% for clean and multi-conditional training respectively which were approximately 24% better than the unadapted HMM.

1. INTRODUCTION

As speech recognition technologies are applied to different applications, their robustness under different channel and environmental conditions becomes more important. Over the past decade, significant progress has been made in robust speech recognition. In [1, 2], front-end processing was modified to remove additive and convolution noise. In [3, 4], the likelihood evaluation criterion and the search process were modified to handle additive or impulsive noise. To allow researchers to compare robustness algorithms with the same data and experimental environments, the Aurora tasks [5] were created. The Aurora 2 and 3 corpora were created for small vocabulary tasks while the Aurora 4 corpus [6] was created for large vocabulary tasks.

One way to improve system robustness is to adapt the recognition model to the noisy environments. Adaptation techniques, such as the Maximum Likelihood Linear Regression (MLLR), which was originally design to adapt to new speakers, can also be applied to handle environmental mis-matches including channel changes and additive noises [7]. In typical speaker adaptation, MLLR would require moderate amount of adaptation data. However, in [8], we showed that adapting HMMs to a new environment via MLLR was effective even when the amount of adaptation data was limited to only one single utterance of around 7 seconds. Alternatively, fast adaptation schemes, such as the Reference Speaker Weighting (RSW) which typically requires less adaptation data, may be more effective in environmental adaptation in such condition.

The environmental adaptation scheme tried in [8] used the HMMs as the acoustic models. Meanwhile, the Polynomial Segment Models (PSMs), which jointly model speech within a segment, were shown to outperform the HMMs in large vocabulary continuous speech recognition (LVCSR) tasks under clean condition [9]. Because the PSMs are more constrained, it is not clear whether they are flexible enough to perform well under noisy and channel mis-matched conditions.

In this paper, we explore the robustness properties of the PSMs under different training conditions: clean and multi-conditional training. The test-sets include sets with additive noise and sometimes from different microphones. Because of the usefulness of adaptation for handling changes in environment, we applied the proposed PSM-based MLLR adaptation [10] and the PSM-based RSW adaptation [11] to handle noisy environments. Experiments were performed to compare their effectiveness under unsupervised adaptation with a single utterance of adaptation data using the ETSI advanced front-end features [1].

The rest of the paper is organized as follows. In Section 2, the basic formulation of the PSM is presented. In Section 3, we present the experimental setup and the HMM baseline results of the Aurora 4 experiments. The Aurora 4 experimental results using the PSMs are presented in Section 4. The results of the PSM-based adaptations are presented in Section 5. Finally, the paper is summarized in Section 6

2. POLYNOMIAL SEGMENT MODEL

The definition and parameter estimation of PSMs were first derived in [12]. For a length N speech segment C, the PSM is defined as,

$$C = Z_N B + E, \tag{1}$$

where C is an $N \times D$ feature matrix for N frames of D dimensional feature vector. Z_N is an $N \times (R+1)$ design matrix for an R^{th} order trajectory model that maps the segments of different durations to a range of 0 to 1, B is an $(R+1) \times D$ parameter model matrix and E is the residue error.

The maximum likelihood estimation of the trajectory parameter matrix B is given by,

$$B = [Z_N' Z_N]^{-1} Z_N' C \tag{2}$$

and the corresponding residue error covariance, Σ, is given by

$$\Sigma = \frac{(C - Z_N B)'(C - Z_N B)}{N}. \tag{3}$$

The triplet ,$\{B, \Sigma, N\}$, can be viewed as the sufficient statistics for C. For a set of segments C_1,...,C_K of model m, the maximum likelihood estimation for the PSM parameter matrix $\hat{B}_m$ and the residue covariance $\hat{\Sigma}_m$ are given by

$$\hat{B}_m = \left[\sum_{k=1}^{K} Z'_{N_k} Z_{N_k}\right]^{-1} \left[\sum_{k=1}^{K} Z'_{N_k} C_k\right] \quad (4)$$

and

$$\hat{\Sigma}_m = \frac{\sum_{k=1}^{K} (C_k - Z_{N_k}\hat{B_m})'(C_k - Z_{N_k}\hat{B}_m)}{\sum_{k=1}^{K} N_k} \quad (5)$$

3. AURORA 4 SETUP AND BASELINE

The standard Aurora 4 tasks use the standard ETSI front-end [13] and the Mississippi State University (ISIP) recognizer as the back-end with four-mixture, cross-word triphone models. Models can be trained with clean data (denoted as clean train) or multi-conditional data (denoted as multi-train). The Aurora 4 tests include 14 test-sets with different noise and channel conditions. In [14], the advanced ETSI front-end [1], which incorporated different noise reduction and channel equalization techniques, was used that resulted in significant performance improvement.

Further performance improvement may be obtained by using a different recognition back-end. In [8], we showed that using the HTK (both front-end and back-end) gave significant improvement compared with using the ISIP recognizer. Our HTK setup was similar to [8] with the exception that the HTK front-end was replaced by the advanced ETSI front-end. The 14 test-sets are grouped into the following four families to simplify discussion and results reporting with the number inside the brackets representing the set numbers defined in the Aurora 4 database.

1. Test-set A : clean data (set 1)
2. Test-set B : noisy data with same channel as training (set 2 to 7)
3. Test-set C : clean data with channel mis-match(set 8)
4. Test-set D : noisy data with channel mis-match (set 9 to 14)

Table 1. *Our Aurora 4 baseline under clean and multi-condition train using the HTK system.*

Group	A	B	C	D	Avg.
Clean Train	12.5%	29.4%	30.9%	44.2%	34.7%
Multi Train	14.1%	23.1%	25.7%	35.9%	28.1%

Table 1 tabulates the baseline performances of our Aurora 4 experiments using the clean- and multi-train training. Compared to the results in [14], our HTK system achieves a 7% and 11% relative improvements on clean- and multi-train respectively. But the comparison is not exact. In addition to differences in recognition back-end, both training and test-set were end-pointed in [14] to remove excess silence in the ISIP system [6]. In our HTK system, only the training data were end-pointed to avoid assumption of end-points in the test data. Also compressed speech features were used in [14] while we used the uncompressed features. Nevertheless, results of our HTK system is comparable with the results in [14] and it will serve as the baseline results in this paper.

4. UNADAPTED PSM EXPERIMENTS

Our PSM system was trained following the same procedure as in [15]. Similar to the HMM system, 4-mixture, cross-word triphone models were used. Each triphone was represented by 3 linear segments. While it is computationally feasible to perform full PSM recognition search [9], we currently do not have a cross-word, PSM-based recognizer. Instead, the PSM recognition was performed by re-scoring HMM-generated lattices. However, optimal segment boundaries were searched using the efficient search algorithm [9] during both training and re-scoring.

Table 2. *Aurora 4 experiment using PSM (4mix and 3 token lattice)*

	Word Error Rates [Relative Improvements %] (%)				
Group	A	B	C	D	Avg.
Clean Train	10.4 [16.3]	27.7 [5.8]	29.7 [3.9]	43.9 [0.7]	33.5 [3.2]
Multi-train	12.6 [10.9]	20.5 [11.3]	24.3 [5.6]	33.6 [6.4]	25.8 [8.1]

Table 2 summarizes the PSM recognition performances using lattice re-scoring under different training conditions. The relative improvements over the HMMs are tabulated inside the square brackets. From the clean-train results in the first row, we note that for the clean test (Group A), the PSMs are about 16% better than the HMMs which is consistent with results in [15]. The performance on the noisy data is not as good and the overall gain is only 3%. Using the Match Pair test, the PSMs are significant better than HMMs in groups A and B but not C or D. One reason may be the differences in lattice qualities between the groups. Table 3 summarizes the HMM lattice quality, showed in the third row, which are the relative improvements of the performance of best in the N-best (second row) over the performance of the first best (first row). As can be observed, the lattice quality decreases as mis-match increases. This affects the performance of the PSM lattice re-scoring. Another possible reason is that because the PSMs are more constrained than HMMs, their distributions are possibly sharper with thinner tails that hurt their performance under mismatched conditions.

Table 3. *Quality of lattice as measured by the word error of the first best and the word error of the top 50 N-best under clean train*

	Word Error Rates (%)			
Test-set	A	B	C	D
First best	12.5	29.4	30.9	44.2
Best of 50 N-best	6.3	19.1	21.5	33.9
Relative Impr.	50%	35%	30%	23%

For multi-condition training (last row of Table 2), the PSMs outperform the HMMs by 10% on test-set A and B. For test-sets in mis-matched channels (C and D), the improvements are about 6% resulting in an average improvement of 8.1% which is significantly higher than the clean-train case. Because the noise and channel conditions are observed in training, they probably are captured by the sharper PSM distributions. Except for group C, the PSMs are significantly better than the HMMs as measure by the Match-pair test.

5. ADAPTATION EXPERIMENTS

We applied unsupervised environmental adaptation to improve recognition performance under noisy and mis-matched channel conditions on Aurora 4. In these experiments, a first pass decoding was performed using the SI models and that decoding output was used to estimated the MLLR transformation or the weights in RSW. The adapted models were then used in a second pass decoding. Because the channel and noise conditions differ for each utterance (based on noise type, channel type and SNR) in Aurora 4, unsupervised adaptation was performed using only one utterance as adaptation data for both the HMMs and the PSMs.

5.1. The MLLR Adaptation

For the MLLR-based environmental adaptation, because non-speech segments contain noise and channel information, a single MLLR transformation, including both speech and non-speech (e.g. silence) sounds, was used instead of using separate transforms which is more common in speaker adaptation tasks. The same transformation structure was applied to both systems.

The PSM-based MLLR adaptation is similar to the one in [10] except that in this experiment, the first pass (SI models) decoding used lattice re-scoring instead of N-Best re-scoring. This lattice was generated from the SI HMMs with 3 recognition tokens.

Table 4. *MLLR adaptation on Aurora 4 using the PSMs and the HMMs*

	Word Error Rates [Relative Improvements %] (%)				
Group	A	B	C	D	Avg.
Clean Train HMM	12.2 [2.4]	28.4 [3.5]	26.7 [13.4]	41.9 [8.5]	32.9 [5]
Clean Train PSM	10.2 [2.1]	26.6 [4.0]	27.3 [8.0]	41.9 [4.4]	32.1 [4.2]
Multi Train HMM	13.7 [2.8]	22.7 [1.7]	24.5 [4.7]	35.3 [1.7]	27.6 [1.8]
Multi Train PSM	12.4 [1.6]	20.3 [1.0]	22.8 [6.2]	33.1 [1.5]	25.4 [1.6]

Table 4 shows the results of the HMM-based and the PSM-based MLLR. Their relative improvements as compared to their corresponding SI models are tabulated inside the square brackets. For the clean test, because there is no channel mismatch nor noise, MLLR is only adapting the speaker with the very limited amount of data (1 utterance) and resulted in only a small gain. This is consistent with what we observed in [8]. Relative improvements on group C and D are better than group B probably because the mismatched channel, which is unobserved in training but is stationary across the test utterance, may be easier to compensate using a single global transformation than the additive noise whose non-linear nature can affect the cepstral coefficients differently depending on the instantaneous SNR. We should point out that the gain from MLLR adaptation is much smaller for multi-train than clean-train. Our conjecture is that under multi-train, the channel and noise conditions are already partially captured by the SI model and this reduces the benefit of the adaptation.

For the PSM-based MLLR adaptation, although the PSMs allow the flexibility of shifting the linear factor [10], it was not used in these experiments. Comparing the relative improvements between the HMM-based MLLR and the PSM-based MLLR, the PSM-MLLR is somewhat less powerful in group C and D but the relative improvement on average, does not differ significantly and the overall accuracy of the PSMs after MLLR still out-perform the HMMs for both clean- and multi-train training.

5.2. The RSW Adaptation

Because in Aurora 4, noise and channel condition changes between utterances, the amount of adaptation data is limited to a single utterance. MLLR, which still requires a fairly high number of parameters to be estimated, may not be the best choice for the speaker or environment adaptation in this case. Instead, RSW adaptation can be used. In RSW, the adapted model is built as a weighed sum of a set of reference speaker models where only the weights are learned from the adaptation data. Thus, the number of free parameters is equal to the number of reference speakers which can be relatively small.

The PSM-based RSW adaptation is described in [11]. In both the HMM and the PSM systems, all the training speaker were used as reference speakers.

Table 5. *RSW adaptation on Aurora 4 using the PSMs and the HMMs*

	Word Error Rate [Relative Improvement %] (%)				
Group	A	B	C	D	Avg.
Clean Train HMM	10.3 [17.0]	23.8 [19.0]	25.5 [17.4]	37.3 [15.6]	28.7 [17.2]
Clean Train PSM	9.7 [7.0]	24.7 [10.8]	25.8 [13.1]	39.2 [10.8]	29.9 [10.7]
Multi Train HMM	12.5 [11.0]	20.1 [13.0]	19.6 [23.9]	30.3 [15.6]	23.9 [15.0]
Multi Train PSM	12.0 [4.5]	19.2 [6.8]	21.0 [13.7]	30.8 [8.3]	23.8 [7.6]

Table 5 shows the results of the RSW adaptation and their relative improvements over the corresponding SI models. We observe that significant improvements were obtained on both the PSM and HMM-based RSW. Compared with the MLLR, the RSW is better able to handle such small amount of adaptation data. The clean test (set A) is an indicator of the portion of gain coming from adapting to the speaker. While the RSW obviously works well for set A, the results of using such a constrained approach for environmental adaptation in other sets are not obvious. Because RSW actually constrains the adapted model parameter space to be spanned by the reference speaker model parameters, this speaker sub-space may not be a good sub-space for noisy or channel mis-matched data especially for clean-train because no noise nor the channel information is available during training. However, since the relative improvement of mis-match test-set is higher than the clean test-set, it shows that the improvement from RSW adaptation comes not only from better adapting to the speaker but also in adapting to the environments. This seems to suggest that with a set of 84 reference speakers from training, the sub-space spanned by these reference speakers does capture a good part of the variations due to channel and noise. For multi-train, it seems that a mix of speakers with different channels and noise actually reduces the gain from speaker adaptation as seem in the reduced gain in the clean test. Similarly to the clean train, the noisy and channel mis-match sets gave better relative improvements.

When we compare the performance of the PSM-based and

the HMM-based RSW performances, we find that the PSM-based RSW did not work as well. One possible reason is again the lattice quality. Since the unadapted HMMs was used to generate the re-scoring lattice, the adapted PSMs may not have a big enough search space to show its improvement. To test this hypothesis, we re-scored the adapted PSMs using the lattices generated from RSW-adapted HMMs. In order to have fair comparisons, the adapted lattice was used only for adapted re-scoring. Table 6 shows the performances of the PSM-based RSW using the adapted HMMs lattice. Significant improvement was shown after replacing the lattice such that the PSM-based RSW now outperformed the HMM-based RSW. This gives our best Aurora-4 performances of 26.5% and 21.3% for clean-train and multi-train respectively. The gain can come from a combination of two factors: the combinational effect of the HMM- and PSM-adapted models or the improved lattice space. However, for clean train, the drastic improvement in sets B, C and D over improvements of set A suggests that the lattice quality is more likely the cause.

Table 6. *The PSM-based RSW on Aurora 4 clean and multi-condition trained models with RSW adapted HMM lattice*

	Word Error Rates (%) [Relative Impr. over unadapted PSMs(%)]				
Group	A	B	C	D	Avg.
Clean Train	9.1 [12.3]	22.0 [20.7]	23.0 [22.6]	34.5 [21.4]	26.5 [21.0]
Multi Train	10.9 [12.0]	17.7 [13.7]	17.0 [30.0]	27.3 [18.7]	21.3 [17.6]

6. CONCLUSION

In this paper, we examined the robustness properties of the PSMs using the Aurora 4 corpus. We found that the improvements of the PSMs over the HMMs were smaller in noisy environments, with the average improvement of 3.2% and 8.1% for clean and multi-condition training respectively. By applying the PSM-based MLLR and RSW adaptation, we showed that the PSM-based RSW not only performed better in capturing speaker characteristics in clean test condition, it is also better in adapting to noise and channel mis-matched conditions.

To obtain the best results, we made use of the adapted HMM lattice for the PSM-based RSW experiments that achieved the word error rates of 26.5% and 21.3% for clean-train and multi-train respectively, which were about 23% relatively better than the unadapted HMMs.

Since the lattice quality can be a limiting factor for the PSMs especially for the noisy tasks such as Aurora 4, we plan to develop a single-pass cross-word PSM-based LVCSR system.

7. ACKNOWLEDGMENT

This work is partially supported by HK Government Research Grant Council grants, CERG grant #HKUST-619505 and CAG grant # CA02/03.EG05.

8. REFERENCES

[1] ETSI, "ETSI es 202 050 v1.1.3 speech processing, transmission and quality aspects (stq); distribution speech recognitionl advanced front-end feature extraction algorithm; compression algorithms," Tech. Rep., 2003.

[2] Y. P. Lai and M. Siu, "Maximum likelihood normailzation for robust speech recognition," in *Eurospeech*, 2003.

[3] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transcations of Acoustics, Speech, and Signal Processing*, vol. 4, pp. 352–359, 1996.

[4] M. Siu and Y. C. Chan, "Robust speech recognition against packet loss," in *Eurospeech*, 2003.

[5] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000: Automatic Speech Recognition: Challenges for the Next Millennium*, 2000.

[6] N. Parihar and J. Picone, "DSR front end LVCSR evaluation au/384/02," 2002.

[7] P. Rahavan, "Speaker and environment adaptation in continuous speech recognition," in *CAIP Technical Report No. TR-227*, 1998.

[8] S.K. Au Yeung and M. Siu, "Improved performance of Aurora 4 using HTK and unsupervised MLLR," in *Proceedings of ICSLP 2004*, 2004, pp. 161–164.

[9] C.F. Li, M. Siu, and S.K. Au Yeung, "Recursive likelihood evaluation and fast search algorithm for polynomial segment model with application to speech recognition," to be appear on IEEE Trans of Speech and Audio Processing.

[10] S. K. Au Yeung and M. Siu, "Maximum likelihood linear regression for sub-phonetic polynomial segment model," submitted to IEEE Signal Processing Letter.

[11] S. K. Au Yeung and M. Siu, "Reference speaker weighting adaptation for sub-phonetic polynomial segment model," in *ICASSP*, 2006.

[12] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," in *Proceedings of ICASSP 93*, 1993, pp. 447–450.

[13] ETSI, "ETSI es 201 108 v1.1.3 speech processing, transmission and quality aspects (stq); distribution speech recognitionl front-end feature extraction algorithm; compression algorithms," Tech. Rep., 2003.

[14] N. Parihar and J. Picone, "Analysis of the Aurora large vocabulary evaluations," in *Eurospeech*, 2003.

[15] S.K. Au Yeung, C.F. Li, and M. Siu, "Sub-phonetic polynomial segment model for large vocabulary continuous speech recognition," in *Proceedings of ICASSP 2005*, 2005, pp. 193–196.