REFERENCE SPEAKER WEIGHTING ADAPTATION FOR SUB-PHONETIC POLYNOMIAL SEGMENT MODELS

Siu-Kei AU YEUNG, and Man-Hung SIU

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

ABSTRACT

Speaker adaptation has been widely used in speech recognition. With small amount of adaptation data, Reference Speaker Weighting (RSW) adaptation was previously proposed for fast HMM adaptation, and has been shown to outperform the more commonly used maximum likelihood linear regression (MLLR) adaptation. Extending our previous work [1, 2] of applying the Polynomial Segment Models (PSMs) in large vocabulary continuous speech recognition (LVCSR) on the WSJ Nov 92 evaluation, we derive the PSM-based RSW fast adaptation technique in this paper. Different from the HMMs, in which the model means are constants within a state, the PSM means are curves represented by polynomials. Experimental results showed that the PSM-based RSW gave approximately the same relative improvement over the unadapted model as in the HMM case. Comparing the PSM-based RSW and MLLR, the PSM-based RSW is more powerful when the amount of adaptation data available is limited. However, it could quickly saturate with increase in adaptation data.

1. INTRODUCTION

Speaker adaptation has been widely used in speech processing such as speaker verification and speech recognition to adjust the parameters of the Speaker Independent (SI) models into the Speaker Adapted (SA) models that match the test speaker. Different adaptation algorithms, such as the Maximum A'Posterior Probability (MAP) [3] adaptation, the Maximum Likelihood Linear Regression (MLLR) adaptation [4] and the EigenVoice (EV) [5] adaptation were proposed for adapting HMMs. The effectiveness of these algorithms depends on the amount of adaptation data.

Reference speaker weighting (RSW) adaptation was first proposed by Hazen and Glass [6] for fast speaker adaptation with limited amount of adaptation data. Similar to EV, RSW estimates a set of weights to combine the reference speaker vectors into a SA model. Because the set of weights contains only a small number of parameters, they can be reliably estimated with a small amount of adaptation data. One difference between EV and RSW is that the speaker space in RSW is simply the set of reference speaker means, instead of the set of orthogonal eigenvectors in EV.

Most adaptation approaches were developed for the HMMs. In recent years, researchers have examined alternatives to the HMMs for representing the speech acoustics. One such alternative is the segment models [7] that are generalizations of the HMMs but explicitly represent the speech dynamics and temporal correlations between frames. The Polynomial Segment Model (PSM) [8] is one type of segment models that represents the speech acoustics with polynomial functions.

Complexity used to be a major limitation of the PSMs. In our previous work [2, 9], we proposed a fast likelihood computation algorithm that significantly improved the PSM recognition and training efficiency and in [1], we applied the PSMs on the large vocabulary recognition tasks.

In order to make the PSM system comparable with the state-of-the-art HMM systems, adaptation algorithms are needed. In [10], PSM-based MAP adaptation was proposed. In this paper, we focus on the task of rapid adaptation and derive the PSM-based RSW algorithm. Because the model means of the PSMs are curves instead of constants (as in the HMMs) and because of the segmental nature of the model, one cannot directly applied the HMM-based RSW derivation for the PSMs.

In addition to deriving the PSM-based RSW adaptation, we compare the PSM-based RSW with the HMM counterpart experimentally using the Wall Street Journal (WSJ0) corpus. We also compare the RSW algorithm with the MLLR under different amounts of adaptation data. These results show that the PSM-based RSW performs similarly to the HMM-based RSW but that RSW is better than MLLR when the amount of adaptation data is small (less than 5 utterances) but MLLR is significantly better for larger amounts of adaptation data.

The organization of this paper is as follows. In Section 2, the basic formulation of the PSMs is presented. In Section 3, we present the proposed PSM-based RSW adaptation algorithm. In Section 4, we report the experimental setup and results using the WSJ0 (standard SI-84 WSJ train-set and Nov'92 5000 words evaluation set). The paper is then concluded in Section 5.

2. POLYNOMIAL SEGMENT MODEL

The PSMs were first proposed in [8]. For a speech segment C with N frames of D dimensional features, the $N \times D$ feature matrix C is modeled as

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

where Z_N is an $N \times (R+1)$ design matrix for an R^{th} order trajectory model that normalizes all segments to unit length, B is an $(R+1) \times D$ parameter model matrix, and E is the residue error.

The maximum likelihood estimation of 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 $\hat{B}_m$ and $\hat{\Sigma}_m$ 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

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

By considering a PSM as a Gaussian distribution with time varying mean and covariance Σ, the log likelihood of the j-th segment, $C_j = O_{\hat{\tau}_j}^{\tau_j}$, with length $N_j = \tau_j - \hat{\tau}_j + 1$, where τ_j is the segment end and $\hat{\tau}_j$ is the segment beginning, can be written as

$$L(O_{\hat{\tau}_j}^{\tau_j} | \hat{B}_m, \hat{\Sigma}_m) = -\frac{N_j}{2}\left[D\log(2\pi) + \log|\hat{\Sigma}_m|\right] - \frac{1}{2}\sum_{t=1}^{N_j}\left[(o_{t+\hat{\tau}_j-1} - b_{t,m})\hat{\Sigma}_m^{-1}(o_{t+\hat{\tau}_j-1} - b_{t,m})'\right] \quad (6)$$

where $b_{t,m}$ is the t^{th} row of the matrix $Z_{N_j}\hat{B}_m$. Because of space limitation, we cannot further describe the PSMs. Readers can refer to [2, 8] for a more in-depth discussion.

3. PSM-BASED RSW ADAPTATION

To improve the PSM recognition performance, speaker adaptation is needed. In [10], the PSM-based MAP adaptation was proposed while in [11], we proposed the PSM-based MLLR adaptation. For very small amount of adaptation data, or for very fast adaptation, EV or RSW adaptations are needed. Because the difference between EV and RSW lies in the representation of the "reference speakers", the adaptation procedure described here can also be applied for the PSM-based eigenvoice adaptation even though we use the RSW notations and terminologies for convenience.

Given a set of speaker reference models, which typically are the speaker models in the training data, the PSM-based RSW finds the set of weights such that the weighted linear combination of the reference speaker models maximizes the likelihood of the adaptation data. Suppose there are S reference speakers. For the m-th Gaussian, denote $B_{si,m}$, $\Sigma_{si,m}$ as the mean and variance of the SI model, $B_{i,m}$ as the mean of the i-th reference speaker, and $\tilde{B}_m$ as the mean of the adapted models. Also denote k_r to be the weight of the r^{th} speaker reference. Then,

$$\tilde{B}_m = \sum_i k_i B_{i,m}. \quad (7)$$

Because the amount of adaptation data is small, the model variances are not adapted.

To maximize the likelihood of the adaptation data and because of the hidden nature of the model "states", the expectation-maximization (EM) algorithm is used that requires only the maximization of the expected log likelihood, $Q(\lambda, \hat{\lambda})$. Define the posterior probability of the segment that ends with model m at time t with duration d as $\gamma_{t,d}(m)$.

$$\gamma_{t,d}(m) = \frac{p(q_{s(t)} = m, \hat{\tau}_{s(t)} = t - d + 1, \tau_{s(t)} = t, O_1^T, |\lambda)}{P(O_1^T)},$$

where $s(t) = m$ denotes the segment that includes time t is from state m. Note that this notation is needed for segment representation because a segment covers multiple frames. Also note that because of the segmental nature, the duration d is one of the indexes in the posterior probabilities. This makes explicit that fact that two segments from the same model and ends at the same time but have different durations are considered different in the PSMs and each has its own posterior probability.

The auxiliary function $Q(\lambda, \hat{\lambda})$ can be written in terms of the posterior probability $\gamma_{t,d}(m)$.

$$Q(\lambda, \hat{\lambda}) = P(O_1^T)\sum_{m,t}\sum_{d=1}^{t}\gamma_{t,d}(m)\log P(O_{t-d+1}^{t}|\tilde{B}_m, \Sigma_{si,m})$$

$$= P(O_1^T)\sum_{m,t}\sum_{d=1}^{t}\gamma_{t,d}(m)\left\{\left[-\frac{d}{2}[D\log(2\pi)] + \log|\Sigma_{si,m}|\right] - \frac{1}{2}\sum_{f=1}^{d}\left[\left(o_{t-d+f} - \sum_{i=1}^{S} k_i\psi_{i,d,m,f}\right)\Sigma_{si,m}^{-1}\left(o_{t-d+f} - \sum_{i=1}^{S} k_i\psi_{i,d,m,f}\right)'\right]\right\} \quad (8)$$

where $\psi_{i,d,m,f}$ is the f^{th} row of $Z_d B_{i,m}$ and is the sampled mean of the i-th reference speaker at the f-th frame from the beginning of the segment. Z_d encapsulates the mapping between the unit length polynomial mean and the length d duration of the segment. Because the PSMs are segment based, summing over all the possible durations is necessary. For each segment, the inner sum over f "accumulates" the frame-by-frame likelihoods.

To find the maximum, we start by differentiating $Q(\lambda, \hat{\lambda})$ with respect to the r-th weight, k_r and setting it to zero. We have

$$0 = \sum_{m,t}\sum_{d=1}^{t}\sum_{f=1}^{d}\gamma_{t,d}(m)\left(o_{t-d+f} - \sum_{i=1}^{S} k_i\psi_{i,d,m,f}\right)\Sigma_{si,m}^{-1}\psi'_{i,d,m,f} = \sum_{m,t}\sum_{d=1}^{t}\sum_{f=1}^{d}\gamma_{t,d}(m)\left(o_{t-d+f} - \sum_{i=1}^{S} k_i\psi_{i,d,m,f}\right)A_{m,r,d,f} \quad (9)$$

where $A_{m,r,d,f} = \Sigma_{si,m}^{-1}\psi'_{r,d,m,f}$ and is known. By re-arranging the terms, we have

$$\sum_{m,t}\sum_{d=1}^{t}\gamma_{t,d}(m)\sum_{f=1}^{d} o_{t-d+f}A_{m,r,d,f} = \sum_{i=1}^{S}\left(\sum_{m,t}\sum_{d=1}^{t}\gamma_{t,d}(m)\sum_{f=1}^{d}\psi_{i,d,m,f}A_{m,r,d,f}\right)k_i \quad (10)$$

Since one such equation can be written for each reference speaker, the S unknown weights can be solved using the S equations by standard techniques for solving system of equations [1].

[1] This is true only if the coefficient matrix is invertible.

Instead of using the posterior probability $\gamma_{t,d}(m)$, one can also simplify the estimation of reference speaker weights by using the Viterbi state alignment. This, in effect, approximates the summation by maximization and makes the γ's into indicator functions.

3.1. Reference Speakers Estimated using MLLR

During training, there may not be enough training data per speaker to estimate the SD models. Instead, one may estimated the reference SD models using MAP or MLLR adaptation. For reference speaker models obtained using MLLR adaptation, Equation 10 can be simplified. Define $W_{r,m}$ to be the MLLR transformation matrix for the m^{th} mixture of speaker r. Then, Equation 10 can be rewritten as

$$\sum_{m,t}\sum_{d=1}^{t}\gamma_{t,d}(m)\sum_{f=1}^{d}o_{t-d+f}\hat{A}_{m,r,d,f} = \sum_{i=1}^{S}\left(\sum_{m,t}\sum_{d=1}^{t}\gamma_{t,d}(m)\sum_{f=1}^{d}\psi_{d,m,f}W_{i,m}\hat{A}_{m,r,d,f}\right)k_i \quad (11)$$

where $\psi_{d,m,f}$ is the f-th row of $Z_d\xi_m$. ξ_m is an $(R+1)\times(D+R+1)$ matrix composed of the original parameter matrix $\hat{B}_{si,m}$ with an $(R+1)$ dimensional identity matrix I_{R+1}. $\hat{A}_{m,r,d,f} = \Sigma_{si,m}^{-1}(\psi_{d,m,f}W_{r,m})'$ which again is known.

While computationally, this is not a big saving, it can significantly reduce memory usage. Instead of storing all the reference speaker models during the adaptation process, it is now only necessary to store the transformation matrices which are only a very small fraction of the parameters to store compared to the full reference speaker models. Note that by expressing the likelihood as a function of $W_{i,m}$ in Eqn 11, different reference speakers are not required to share the same regression class definitions.

4. EXPERIMENTS

LVCSR experiments were performed on the ARPA Wall Street Journal (WSJ) 5k word task [12] with models trained using the standard SI-84 train set (7138 utterances) and tested on the Nov'92 5000 word evaluation set (330 utterances). The HMM training and decoding procedure and settings were similar to [1]. In short, we used crossword triphone models with 16 mixture components, tied with a decision-tree based clustering that results in approximately 3000 tied-states. With the standard bigram language model, our best HMM baseline achieved a 7.81% WER which is comparable with results from other researchers using the same test-set and conditions [13][2]

For the PSMs, each phoneme was represented by 3 independent sub-phonetic segments which can be viewed as a special case of the dynamic multi-region PSM [2]. Because of using 3 segments per phoneme, only first order PSM (linear) was used instead of the more commonly used second order (quadratic) PSMs. The PSM training followed the procedure described in [9]. The 8-mixture SI PSMs were trained with mixture splitting generally followed [1]. However, instead of using the HMM state-tying information for the parameter sharing, we applied the PSM-based decision tree tying such that the PSM sub-phonetic segments were treated like a HMM state resulting in approximately 3000 tied states. Our previous results showed that the PSM-based tree outperformed the HMM-based tree when large number of mixtures (4 or above) were used. The PSM recognition was performed using N-Best and lattice re-scoring in which the N-best and lattices were generated using the SI HMMs. While it is possible to perform a full PSM search, our current PSM implementation does not support cross-word triphone decoding. Our re-scoring, however is different from other re-scoring work [14] in that the HMM alignment was not used. Instead, a full search for optimal segment boundaries was performed using the fast PSM computation [2]. In this paper, a 3-token lattice was generated from the 8-mixture HMM with the N-Best size of 50. Our PSM 8-mixture SI result of 6.86% word error rate is 12.2% relatively better than the corresponding 16 mixture HMM results but used about 25% less parameters.

4.1. RSW adaptation

The HMM-based RSW adaptation was performed on the 16-mixture SI HMMs. All the available speakers in training (about 80) were used as reference speakers[3]. The SD models for the reference speakers were built using MLLR adaptation with 32 regression classes [4] which were shared by all reference speakers. The RSW adaptation was unsupervised and mostly with single adaptation utterance. Each utterance was first decoded using the SI model and then, the RSW adaptation was carried out using decoding output. After determining the most likely weights to create the SA models, the utterance was decoded again using the adapted models. For the PSMs, the adapted models were used to re-score the lattice generated by the unadapted (SI) HMMs.

Model	WER	Relative Imp.
HMM (SI, 16mix)	7.81%	-
HMM (RSW, 1utt)	7.0%	10.3%

Table 1. *Performance of the HMM-based RSW adaptation*

Tables 1 and 2 tabulate the results of the HMM-based and the PSM-based RSW adaptation respectively with only one adaptation utterance which is about 7 seconds on average. For the HMMs, the RSW adaptation gave about 10% relative improvement while for the PSMs, the improvement was about 7% relative. Both results show that RSW is useful for fast adaptation when only very limited amount of adaptation data is available. The relative improvement of the PSM-based RSW is worse than the HMM-based RSW adaptation. One should note that our adapted PSMs are in fact about 10% relative better than the HMMs in the overall word error rate and thus, improvement from RSW may be smaller. Another possible reason is in the lattice quality used in the re-scoring paradigm. These lattices were generated by the unadapted HMMs which have a one-best performance of almost 1.5% worse than the adapted PSMs. To see whether lattice quality is an issue, lattices generated by the adapted HMMs (with WER 7.0%) were used and rescored by the RSW adapted PSMs. These results are shown in the third row of Table 2 which gave a small improvement. This improvement can come from the combination of two possible sources: a

[2]Better results may be achievable using a better lexicon, more training data or a trigram language model. Given our limited resource, we used the standard setup for ease of comparison

[3]While slightly better performance may be obtained by careful selection of reference speakers, it is not the focus of this paper and using all speakers allows us to avoid the heuristic of reference speaker selection.

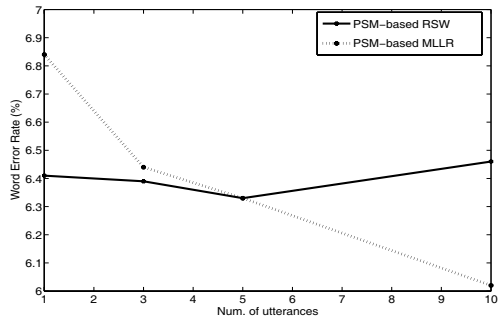

Fig. 1. Comparison between the PSM-based RSW and MLLR

better search space in the lattice or, the combinational effect of both the HMM adapted model and the PSM adapted model.

Model	WER	Relative Imp.
PSM (SI, 8mix)	6.86%	-
PSM (RSW, 1utt)	6.41%	7%
PSM (RSW, adapted lattice)	6.3%	8.1%

Table 2. *Performance of the PSM-based RSW adaptation*

4.2. Comparison between the PSM-based RSW and MLLR adaptation

While the RSW adaptation works very well with very limited adaptation data, it is interesting to see how it performs compared to MLLR under different amounts of adaptation data, and find the point at which these two algorithms intersect.

Figure 1 shows the performance comparison between the PSM-based MLLR and the PSM-based RSW adaptation against various amounts of adaptation data (shown by the number of utterances). When the amount of adaptation data was very limited such as with only one utterance, the RSW performed significantly better than MLLR. With such small amount of adaptation data, the MLLR adaptation is not very useful. When the amount of adaptation data increases, MLLR consistently improves while the RSW improves slightly up to 5 utterances. The results for RSW are so flat that between 1 to 10 utterances of adaptation data, the performance variation is less that 0.1% absolute error. This is consistent with our intuition that because the number of adaptation parameters is small for RSW, it is only good for very limited amount of adaptation data.

5. CONCLUSION

In this paper, we extended the PSMs by deriving the PSM-based RSW to improve the recognition accuracy. We observed that the proposed PSM-based RSW adaptation gave similar, but slightly worse relative improvement compared to the HMM-based RSW adaptation. Possible reasons include the relatively better SI performance of the PSMs and the quality of the lattice in re-scoring. In order to resolve the lattice quality problem, one of our future work in PSMs is to implement a cross-word decoder to perform full recognition search.

In addition, we also showed that the PSM-based RSW adaptation was only suitable for fast adaptation with very few adaptation utterances (say a few seconds). When more data are available, other adaptation algorithms like the MLLR, should be applied.

6. ACKNOWLEDGMENT

This work is partially supported by HK Government Research Grant Council CERG grant #HKUST/619505 and CAG grant 02/03.EG05.

7. REFERENCES

[1] 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.

[2] C.F. Li and M. Siu, "Training for polynomial segment model using the expectation maximization algorithm," in *Proceedings of ICASSP 2004*, 2004, pp. 841–844.

[3] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation for Markov chain," *IEEE Trans. on Speech and Audio Processing*, pp. 291–298, 1994.

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

[5] R. Kuhn, J.C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on Speech and Audio Processing*, pp. 695–707, 2000.

[6] T. J. Hazen, "A comparsion of novel techniques for rapid speaker adaptation," *Speech Communication*, vol. 31, pp. 15–33, 2000.

[7] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From hmm's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans of Speech and Audio Processing*, vol. 4, pp. 360–387, 1996.

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

[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] A. Kannan and M. Ostendorf, "Adaptation of polynomial trajectory segment models for large vocabulary speech recognition," in *Proceedings of ICASSP 1997*, 1997, pp. 1411–1414.

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

[12] D. Paul and J. Baker, "The design of Wall Street Journal-based csr corpus," in *Proceedings of ICSLP 1992*, 1992, pp. 899–902.

[13] D. Pallett, J. Fiscus, W. Fisher, and J. Garofolo, "Benchmark tests for spoken language program," in *DARPA Human Language Technology WorkShoop*, 1993.

[14] M. Siu, R. Iyer, H. Gish, and C. Quillen, "Parametric trajectory mixtures for lvcsr," in *Proc. of ICSLP*, 1998.