A Low Complexity Cluster Model Interpolation based On-Line Adaptation Technique for Spoken Query Systems

S Shahnawazuddin and Rohit Sinha

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati -781039, India
{s.syed, rsinha}@iitg.ernet.in

Abstract

The work presented in this paper describes the issues of on-line adaption in context of spoken query systems. In such systems, the available adaptation data is extremely small ($\leq$ 3 seconds). Consequently, adapting such systems becomes extremely challenging. Moreover, since these systems are meant for real-time applications, the employed adaptation technique should not add much latency to the system response. To address these issues, a simple cluster model interpolation based approach for on-line adaptation is presented in this work. The proposed approach employs an OMP based search scheme to select a set of acoustically close models from a set of pre-trained cluster models. The selected cluster models are then linearly interpolated to derive the adapted model parameters. In this work, these interpolation weights are derived from the sparse coefficients in an approximate manner. Such an approximate approach helps in avoiding the iterative ML weight estimation usually employed in existing techniques. The proposed adaptation approach though not optimal, is found to be effective for on-line adaptation. The same has been verified in this work for an LVCSR task and also for an Assamese name recognition system which is a typical example of such query systems.

Index Terms: Spoken query system, fast adaptation, on-line adaptation, acoustic model interpolation, sparse representation.

1. Introduction

In the past few decades the application of speech processing has spread to a great majority of areas. One such application is the automatic speech recognition (ASR) based interactive voice response (IVR) system [1]. With the advancements made in speech recognition research, the touch tones that were used earlier have been done away with in the present IVR systems. Such systems, referred to as *spoken query* (SQ) system or *spoken dialogue* (SD) system, rely on a hidden Morkov model (HMM) based ASR system to decode the user input and then disseminate the desired information [2, 3]. A similar system has been developed recently for the Assamese agricultural commodity name recognition and price information dissemination [4]. The developed SQ system enables the user to make a query about the price of any agricultural commodity over telephone network (land-line/mobile). The spoken query is recorded and processed and then the current price of that commodity (in a desired district) is disseminated through pre-recorded voice responses.

In the work presented in this paper, we have explored the issues of adapting SQ systems to the end user in order to enhance their performance. In case of such systems, as also pointed in [3], the available data for adaptation is very small ($\leq$ 3 seconds) as the user inputs are either isolated words or small sentences only. The caller may input successive queries during a single call but even then it may not last for more than 10-15 seconds in the best cases. Moreover, the system is always blind towards the identity of the user and has to treat each call independent of all previous calls. Consequently, adapting the system to the end user in incremental mode also becomes infeasible. In addition to that, since these systems are meant for real-time applications, the latency in the system response is also a major factor while performing adaptation. Under these conditions, the conventional techniques like [5, 6, 7] etc, become ineffective as they are required to estimate a large number of parameters and the available data is not sufficient for the same. The fast adaptation techniques like [8, 9, 10, 11] etc, on the other hand, have been explored in the recent past to perform adaptation under such low data scenario. In this paper, some of the fast adaptation techniques have been explored to perform on-line adaptation of a SQ system. These technique generally assume that the adapted model parameters (Gaussian means and/or mixture-weights) lie in a low dimensional space spanned by a set of bases (predefined acoustic models). Such approaches are found to be very effective in low data conditions since only a small number of parameters are required to be estimated.

In case of SQ query systems, the effectiveness of the aforementioned techniques gets diluted since no supervision in terms of the true transcription of the adaptation data is available. Moreover, transform parameters are required to be estimated for each test utterance independently since speaker specific tying is not possible. In this work, apart from exploring the existing model interpolation based fast adaptation approaches for on-line tasks, a low complexity adaptation approach to overcome the aforementioned issues is also proposed. In the proposed approach, the adapted model parameters are derived by a dynamic selection of pre-trained acoustic cluster models and their linear interpolation in desired parameter space. In order reduce the complexity involved in the estimation of interpolation weights, these are derived in an approximate manner. Such intuitive approximations help in keeping the system latency low. When evaluated on WSJCAM0 database, the proposed approach is found to result in performances slightly inferior to that of the existing similar techniques despite the use of approximate weights. Motivated by these results, the proposed approach is also implemented on a Assamese name recognition system. In this case, the available data is only 1-2 seconds in duration. The proposed technique is found to be effective even under such an extremely low data adaptation scenario.

The rest of this paper is organized as follows: The proposed low complexity on-line adaptation approach is detailed in Section 2. The experimental evaluation of the proposed technique is given in Section 3. The paper is finally concluded in Section 4.

978-1-4799-4219-0/14/$31.00 ©2014 IEEE

2. Proposed low complexity approach

As already discussed, in case of SQ systems, the system has to respond to the end user with a minimal latency. Consequently, apart from the effectiveness of the employed adaptation technique in low data conditions, the computational complexity also plays a major role. In case of model interpolation based adaptation approaches, there are two major factors that add to the latency in implementation, namely:

(i) Selecting a set of acoustically close bases from a pool of candidate models to be interpolated.

(ii) A robust estimation of the interpolation weights corresponding to the selected bases.

In the following we discuss the explored approaches to minimize the latency due to both of the aforementioned factors.

2.1. Cluster model selection scheme

The model interpolation based adaptation techniques employ a dynamic selection of acoustically close bases from a pool of candidate acoustic models. These candidate acoustic models can be defined either by creating a speaker adapted (SA) model for each of the speakers in the training set as done in [8, 10, 11] or by clustering the speakers using some similarity criteria [9]. Employing SA models is reported to be very effective as it provides a greater amount of acoustic and linguistic diversity by capturing the intra- and inter-speaker variability. On the other hand, such an approach results in increased search complexity due to a large number of candidate bases. In case of SQ systems, this can be reduced by clustering the speakers in the training set into a smaller number of cluster. This leads to a loss in the finer acoustic details (the intra-speaker variability) due to the pooling of data from a large number of speakers. Deriving such averaged models is the price paid to avoid the increased latency incurred in case of the former approach.

For acoustic clustering, each speaker in the training set is first represented as a supervector derived by concatenating the Gaussian mean parameters of their respective SA models. In this work, MAP adaptation of the Gaussian mean parameter of the monophone HMMs (learned using the entire training data) is done to create the speaker specific supervectors. This ensures that almost all the Gaussians get adapted using the available speaker specific data and hence the derived supervector uniquely represents a speaker [8]. These supervectors are then grouped into a desired number of clusters using vector-quantization (VQ). Pooling the speech data corresponding to all the speakers assigned to a particular cluster, a cluster model is then created using MAP adaptation of the speaker independent (SI) model (triphone HMM). An added advantage of acoustic clustering is that it also reduces the memory requirements for storing the candidate models. Using SA models corresponding to each of the speakers also hampers the system portability [3].

Once the candidate models are created, the next step is to select a set of acoustically close models for each test speaker/utterance and interpolate them. The dynamic selection of acoustic models can be done using the approaches reported in [10, 11] but these are found to be ineffective for on-line tasks. In [10], a set of SA models is selected by doing a Viterbi-alignment based maximum likelihood (ML) search over the SA models corresponding to all the speakers in the training data. In practical cases, the number of speakers in the training set is generally quite large. Consequently, such an approach is very prohibitive for on-line adaptation. In [11], the support speaker vectors are searched through a two class support vector machine (SVM) training with respect to the test speaker/utterance representation. In case of on-line adaptation task, one of the class happens to have only one example (the supervector for the current test utterance). Consequently, such a class imbalance leads to a poor selection of models to be interpolated.

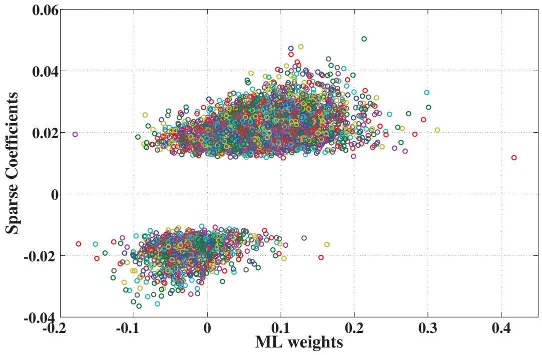

Figure 1: Mapping of the ML weights and the corresponding sparse coefficients for the selected atoms (for all the test utterances). It is quite apparent that the sparse coefficients have a much lower value (a factor of 10) in comparison to the corresponding weights obtained in ML sense.

In this work, we have resorted to the basis selection scheme described in [12]. In that work, sparse representation (SR) techniques [13] are used to select a set of acoustically close models. The selected models are then linearly interpolated using weights estimated in ML sense. This approach is reported to be comparatively less latent than the ML based approach since for selecting K bases from N candidate models, only a single Viterbi based alignment is required. To do the same, first a dictionary is created using the Gaussian mean supervectors of the cluster models as the atoms. Similarly, for each of the test utterance, a *target* supervector is created by extracting the Gaussian means of the adapted model (derived by MAP adaptation of the SI model parameters under the constraints of the first-pass hypothesis). A set of cluster models is then selected for each test utterance using orthogonal matching pursuit (OMP) [14] and is linearly interpolated to derive the adapted model parameters. In order to further reduce the latency, we propose a low complexity scheme to derive the weights in an approximate manner as discussed in the next.

2.2. The approximate interpolation weights

The bases selection approach reported in [12], resorts to the use of the SR techniques like OMP and least absolute shrinkage and selection operator (Lasso) [15]. In that work, only the indices of the atoms corresponding to non-zero sparse coefficients are made use of for model interpolation while the sparse coefficients are discarded. The reason for the same is that the sparse coefficients have a lower dynamic range than that of the corresponding ML weights estimated for the selected bases as shown in Figure 1. This happens due to the normalization of the atoms that is required for the computation of correlation during OMP/Lasso based search. Consequently, the sparse coded target also has a lower dynamic range than that of the dictionary atoms. Such a supervector cannot be used as the Gaussian mean parameters in the adapted acoustic model. This is so because the dynamic ranges of the remaining parameters (covariance, mixture-weights, etc.) are different as they had been jointly estimated with the unnormalized SI mean parameter.

The estimation of interpolation weights in ML sense does helps in getting significant improvements but at the cost increased complexity (an absolute improvement of 0.45% over

Algorithm 1 Proposed low complexity adaptation approach

```
Given: The exemplar dictionary $\tilde{\mathbf{A}}$ obtained by normalizing each of the dimensions of the columns of the matrix of cluster-specific mean-supervectors $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_N]$ with its standard deviation from SI model, $\{\mathbf{a}_i\}_{i=1}^{N}$ are cluster mean supervectors
    Dynamic selection of cluster models
Given: The number of bases to be selected $K$
Step 1: Obtain first-pass hypothesis for test data using SI model
Step 2: MAP adapt the SI model with test data using the first-pass hypothesis with relevance factor ($\tau$) set to zero
Step 3: To form target $\mathbf{b}$, extract mean-supervector, normalize that with SI standard deviation and substitute zeros for all unadapted dimensions
Step 4: Using OMP, sparse code the target $\mathbf{b}$ over the dictionary $\tilde{\mathbf{A}}$ with sparsity $K$
    Deriving approximate weights and adapted model parameters
Given: The array of sparse coefficients, $\{y_i\}_{i=1}^{K}$, obtained by minimizing the representation error after selecting the $K^{th}$ atom
Step 5: For model interpolation, choose the columns from unnormalized matrix $\mathbf{A}$ corresponding to the indices of the non-zero sparse coefficients
Step 6: For $K$ basis models, estimate of the interpolation weights $\{w_i\}_{i=1}^{K}$, with $0 \leq w_i \leq 1$ and $\sum_{i=1}^{K} w_i = 1$, as
    $w_i = \frac{y_i - min\{y_i\}}{\sum_{i=1}^{K} [y_i - min\{y_i\}]}$
Step 7: The adapted model parameter is derived as
    $\hat{\boldsymbol{\lambda}} = \sum_{i=1}^{K} w_i \, \boldsymbol{\lambda}_i$
```

the existing techniques is reported in [12]). In order to reduce the same, we derived approximate interpolation weights from the sparse coefficients. As evident from Figure 1, for majority of the cases, whenever the sparse coefficients are positive, the corresponding ML weights are also positive and vice-versa. Furthermore, if the bases are arranged in the order of the magnitude of the sparse coefficients and ML weights, a very similar ordering is observed in both the cases. The only difference among the two is in terms of a scaling factor. To overcome this, one can normalize these sparse coefficients so that they turn out to be *non-negative* and *sum to one*. This transformation ensures that dynamic range of each of the dimensions in the sparse codded target is similar to that in the dictionary atoms. Moreover, this does not incur any extra computation after bases search unlike the iterative ML estimation process. An added advantage is that these weights can be used for the interpolation of Gaussian mixture-weights as well. In essence, in this manner we try to move the acoustic space towards the test speaker/utterance in proportion to the sparse coefficients with mean interpolation and try to give lower weight to less probable Gaussians among the clusters with mixture-weight interpolation. The mixture-weight interpolation helps in getting extra improvements without any added computation. The proposed on-line adaptation approach is detailed in Algorithm 1. The steps for model selection are the same as in [12] and have been reproduced here for the sake of completeness.

3. Experimental setup and results

To evaluate the performance of the proposed technique on an LVCSR task, an ASR system is developed using HTK [16] on WSJCAM0 corpus [17]. In this database, the training set consists of 7861 utterances from 92 speakers with approximately 90 sentences per speaker. The test set consists of 368 utterances from 20 speakers. Speech is analyzed into 20 ms frames with 10 ms shift and parameterized into MFCC features comprising of 13 static features along with their first and second derivatives. A 3-state left-to-right HMM architecture is used and state-clustered cross-word triphones are trained. Each triphone HMM is modeled using 8 Gaussian mixtures per state. A 5k-bigram language model is used in decoding. OMP-Box v-10 toolkit [18] is used for implementing OMP in Matlab.

Table 1: Performances for the proposed on-line adaptation approach along with those of the existing adaptation methods in utterance-specific mode on WSJ0CAM database. For cluster model interpolation, 4 bases are selected from a set of 8 cluster models.

Adaptation Technique	WER (in %)
Unadapted SI	11.30
MAP (mean-only)	11.25
Global MLLR	11.13
Fixed cluster (8)	10.90
EV (16 bases)	10.63
Im-RSW (18 bases)	10.57
ML search, ML weights	10.70
OMP search, ML weights	10.71
OMP search, prop. weights	10.73

For the adaptation experiments, a dictionary $\mathbf{A}$ is created as explained in Section 2.1 . For bases selection, the target supervector is derived in a similar manner. The relevance factor, τ, is set to zero for the creation of target as well as the dictionary atoms [12]. In this work, all adaptation experiments are performed in the unsupervised utterance-specific (on-line) mode, i.e., the available adaptation data is the current test utterance only. In addition to that, bases selection as well as weight estimation is done under the constraints of the first-pass hypothesis generated using the SI model. In case of MLLR, global transforms are generated since the adaptation data is too low for the regression class based tying to be effective. For the proposed technique, 4 basis models are selected from a set of 8 predefined cluster models. For the proposed approach, both the Gaussian mean and the mixture-weight parameters are adapted to derive the cluster models to be interpolated. In case of ML weight estimation, only Gaussian means are interpolated while in case of the proposed weights, both the Gaussian mean and the mixture-weights are interpolated.

To evaluate the effectiveness of the proposed technique, three different combinations of bases selection and interpolation are tried, i.e bases search using ML/OMP with weights estimated in ML/approximate sense. The performances for those experiments are given in Table 1. The performances for some of the existing techniques are also given in Table 1. It is to note that the MAP and MLLR techniques fail to result in any significant improvements in the system performance. This is so because these techniques estimate a large number of model parameters and the given adaptation data is too low for a robust estimation of so many parameters. Though the performance of Im-RSW [10] is better than all the cluster model interpolation based approaches, it comes at the cost of increased search complexity. Furthermore, both eigenvoices (EV) [19] and Im-RSW require interpolation of a large number of models (16 and 18, respectively). The complexity of the ML weight process increases with an increase in the number of bases being interpolated and hence has a limiting effect in on-line tasks. This point is argued in detail later in the paper. It is evident from Table 1 that the dynamic selection and interpolation of cluster models performs better than the fixed cluster approach (CAT). Furthermore, it is interesting to note that the use of proposed approximate weights also results in a very similar performance. As already discussed, such approximations help in avoiding the latency incurred due to the iterative ML estimation process which is much desired in on-line applications.

3.1. Evaluation on Assamese SQ system

Recently a telephone-speech based inquiry system has been developed for the dissemination of the prices of agricultural commodities to the user in Assamese language. The SQ system employs an ASR system that is developed using HTK [16] on 30 hours of speech data collected from 885 speakers. Speech is analyzed into 20ms Hamming windowed frames with a shift of 10ms and is parameterized into MFCC features comprising of 13 base features with its first and second derivatives.The performance of the system is evaluated on a test set comprising of 2552 isolated word utterances from 275 speaker using an equilikely wordnet and a dictionary of 143 words. Further details about the data collection and that of the overall system performance are available in detail in [4].

The developed SQ system is designed for a wide public usage and the typical user response consists of the commodity name for which the price is inquired. So neither the speaker information nor a large amount of adaptation data is available in this case. Furthermore, being an on-line system, it is also desired to keep the latency low. The proposed approach is hence found to be very promising in its context. As already mentioned, clustering of the acoustic space along with the SR bases search scheme helps in reducing the latency involved in the selection of the bases. In addition to that, the use of approximate interpolation weights helps in avoiding the iterative ML estimation process. Table 2 enlists the performance of the proposed approach along with that for some of the existing techniques. It is to note that the performances for the existing techniques are same as that for the unadapted SI system. This is so because the available adaptation data, i.e., the current test utterance is of the order of 1-2 seconds only in duration since the user responses are in the form of isolated words. The performance of decoding the test utterance using the most likely cluster model is also shown in Table 2. It is interesting to note that the proposed approach outperforms all the discussed adaptation techniques.

3.2. Discussion on computational complexity

All adaptation experiments reported in this work are performed in the utterance-specific mode under the constraints of the first-pass hypothesis. Consequently, the latency incurred due to the the first-pass decoding using the SI system happens to be a common factor across all the discussed techniques. Furthermore, as already discussed, the complexity of the ML weight estimation process depends on the number of models being interpolated. In case of EV and Im-RSW, the best case performances are obtained by interpolating a comparatively larger number of bases as compared to the cluster based approaches. The runtime for estimating 16 weights turns out to be 51 seconds for the EV experiment performed on the WSJCAM0 task[1]. On the other hand, ML estimation takes 18 seconds for deriving 4 weight parameters for the cluster based experiments on the same database. This shows how the clustering of the acoustic space helps in reducing the latency. In addition to that, the proposed approximate weight derivation technique helps in avoiding the latency incurred due to the iterative ML approach. Since the adaptation experiments are unsupervised, a single iteration of weight estimation does not suffice. For the reported experiments, it was observed that on an average 6-7 iterations were required for convergence (4 iterations are reported in [9]). In case of the cluster model interpolation based experiments, the obtained WER was 11.04% only for a single pass weight estimation . The proposed approach, on the other hand, results in an absolute improvement of 0.3% over the single iteration ML estimation case. For the Assamese SQ system, the total processing time for a user query (including the two decodings) turns out to be 19.5 seconds which is very much tolerable.

Table 2: Performances of the proposed adaptation approach for the Assamese commodity name recognition task with 4 bases derived out of 8 acoustic clusters and interpolated in chosen parameter space. Also shown are the performances for the existing adaptation approaches.

Adaptation Techniques		WER (%)
Unadapted SI		16.00
Conv. Adapt. Tech.	Mean only MAP	16.00
	Mean & mix-wt. MAP	16.00
	Global MLLR	16.00
	Global CMLLR	16.00
Max likelihood cluster search		15.00
SR search	Mean interpolation	14.50
	Mix-wt. interpolation	14.60
	Mean & mix-wt interpolation	**14.10**

4. Conclusion

In this work, a cluster model interpolation based adaptation approach is explored in context of spoken query systems. The proposed approach explores the use of SR technique (OMP) for the selection of acoustically close cluster models from a set of 8 pre-trained cluster models. Furthermore, this work presents a low complexity scheme for deriving the interpolation weights from the sparse coefficients. These approximate weights are found to result in improvements very similar to the other existing techniques. The use of such approximate weights helps in keeping the system latency significantly low. The proposed approach is evaluated for performance on a LVCSR task and an Assamese SQ system and is found to be effective for both the tasks. Relative improvements of 5.4% and 12% over the unadapted SI system are obtained for the LVCSR and Assamese name recognition tasks, respectively. The obtained reduction in WER is much higher for the case of the SQ system than that for the LVSCR task. This is due to a large acoustic mismatch between the training and testing speakers in case of the Assamese name recognition task.

[1] Computed using 64-bit HTK (ver. 3.4.1) running on Intel Xeon 6-core CPU, 2.4 GHz with 16 GB RAM.

5. References

[1] L. Rabiner, “Applications of voice processing to telecommunications,” in *Proc. IEEE*, vol. 82, 1994, pp. 199–228.

[2] A. Trihandoyo, A. Belloum, and K. M. Hou, “A real-time speech recognition architecture for a multi-channel interactive voice response system,” in *Proc. ICASSP*, vol. 4, May 1995, pp. 2687–2690.

[3] J. R. Glass, “Challanges for spoken dialogue systems,” in *Proc. IEEE ASRU workshop*, 1999.

[4] S. Shahnawazuddin, D. Thotappa, B. D. Sarma, A. Deka, S. R. M. Prasanna, and R. Sinha, “Assamese spoken query system to access the price of agricultural commodities,” in *Proc. 19th National Conference on Communiaction*, New Delhi, Februray 2013.

[5] J. L. Gauvain and C. H. Lee, “Maximum a-posteriori estimation for multivariate gaussian mixture observations of Markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[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*, vol. 9, pp. 171–185, 1995.

[7] V. Digalakis, D. Rtischev, and L. Neumeyer, “Speaker adaptation using constrained estimation of gaussian mixtures,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, 1995.

[8] T. J. Hazen and J. R. Glass, “A comparison of novel techniques for instantaneous speaker adaptation,” in *Proc. of European Conference on Speech Communication and Technology*, 1997, pp. 2047–2050.

[9] M. J. F. Gales, “Cluster adaptive training of hidden Markov models,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, July 1999.

[10] B. Mak, T. Lai, and R. Hsiao, “Improving reference speaker weighting adaptation by the use of maximum-likelihood reference speakers,” in *Proc. ICASSP*, vol. 1, May 2006.

[11] T. Cai and J. Zhu, “A novel method for rapid speaker adaptation based on support speaker weighting,” in *Proc. ICASSP*, 2005, pp. 993–996.

[12] S. Shahnawazuddin and R. Sinha, “Improved bases selection in acoustic model interpolation for fast on-line adaptation,” *IEEE Signal Processing Letters*, vol. 21, no. 4, pp. 493–497, April 2014.

[13] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer New-York, 2010.

[14] Y. Pati, R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition,” in *Proc.* 27^{th} *Asilomar Conference on Signals, Systems and Computers*, vol. 1, Nov 1993, pp. 40–44.

[15] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[16] The HTK Toolkit: http://htk.eng.cam.ac.uk.

[17] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, “WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition,” in *Proc. ICASSP*, vol. 1, May 1995, pp. 81–84.

[18] OMP-Box v10: http://www.cs.technion.ac.il/∼ronrubin/software.html.

[19] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, “Rapid speaker adaptation in eigenvoice space,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.