A Low Complexity Model Adaptation Approach involving Sparse Coding over Multiple Dictionaries

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 a novel on-line adaptation approach for extremely low adaptation data scenario. The proposed approach extends a similar redundant dictionary based approach reported recently in literature. In this work, the orthogonal matching pursuit (OMP) algorithm is used for bases selection instead of the matching pursuit (MP). This helps in avoiding the selection of an atom more than once. Furthermore, this work also explores the use of cluster-specific eigenvoices to capture local acoustic details unlike the conventional eigenvoices technique. These approaches are then combined to reduce the number of weight parameters being estimated for deriving adapted model. Towards this purpose, separate sparse coding of the test data is performed over a set of dictionaries. Those sparse coded supervectors are then scaled and used as the Gaussian mean parameter in the adapted model. Consequently, only a few scaling factors are needed to be estimated. Such a reduction in number of parameters is highly desirable for on-line applications where the latency is a major factor.

Index Terms: Fast adaptation, on-line adaptation, eigenvoices, redundant dictionaries, sparse representation.

1. Introduction

On-line speech recognition systems, such as spoken query systems [1, 2, 3], are designed to produce the desired response after decoding a spoken query. In such systems, the user queries are either single words or small sentences comprising of three to five words only. These systems typically involve HMM based acoustic/linguistic modeling and hence adapting such systems to the end user becomes a challenging task when the available data is extremely low. In the conventional adaptation techniques like [4, 5, 6], a large number of parameters are estimated. This, in turn, requires a significantly large amount of adaptation data and hence those become ineffective in low data scenario. The fast adaptation techniques like [7, 8, 9, 10, 11] are found to perform comparatively better under such conditions. These techniques assume that the adapted model parameters lie in a low dimensional space spanned by a set of basis vectors. This leads to a huge reduction in the number of parameters to be estimated. Hence, imparting the effectiveness to these techniques for extremely low data scenario. These techniques mainly differ among themselves in the way the bases are determined. Deriving speaker adapted (SA) models from the speaker independent (SI) system by MAP/MLLR adaptation is a way to construct the bases [7, 10, 11]. In the eigenvoices (EV) technique [8], the basis models are obtained by choosing the top $M$ eigenvectors through principal component analysis (PCA) performed on the covariance/correlation matrix constructed out of the adapted mean supervectors of the training speakers. Though EV technique is reported to be efficient in extremely low data conditions, the performance is found to saturate very fast.

In the past decade, a number of techniques have been proposed in order to address the problem of saturation in performance in case of EV even when sufficient amount of adaptation data is available. In [12], the supervectors derived from the speaker adapted (SA) models are segmented into $N$ supervectors before PCA is performed. In the work reported in [13], 2-dimensional PCA has been implemented for creating the basis models. A very recent approach for deriving the basis models involves probabilistic linear discriminant analysis (PLDA) [14]. All of these approaches do perform better than the EV but at the cost of an increase in the number of parameters to be estimated. This, in turn, requires a greater amount of adaptation data for their robust estimation. It is worth noting that for very small amount of adaptation data ($\leq 6$ secs), the reported performances are even below that for the SI system. Another issue with the EV approach and its variants is that the derived eigenvectors cater to the inter-speaker variabilities only. To overcome this issue, a compressive sensing based adaptation technique is proposed in [15]. In that work, a redundant dictionary is employed with supervectors derived from SA models and their corresponding eigenvoices constituting the atoms. This is done in order to capture the intra-speaker variations through the SA models. Matching pursuit (MP) [16] based sparse representation (SR) and $\ell_1$-regularized SR [17] are the employed bases selection techniques. The performances reported show that $\ell_1$-regularization is slightly superior than MP but, at the same time, MP is observed to be much faster than $\ell_1$-regularized SR.

In the work presented in this paper, we have explored the use of dictionary based approach for on-line adaptation. The main contributions of the work are as follows:

i. An OMP based bases search approach over a redundant dictionary is explored which avoids repeated selection of an atom unlike that encountered in the existing MP based bases search approach for fast adaptation [15].

ii. In the EV approach the intra-speaker variations are not captured well. To address this issue, cluster-specific eigenvoices are explored for a comparatively better capturing of finer acoustic details.

iii. In the model interpolation based approaches, the complexity of weight estimation process becomes prohibitive for the on-line applications. To address the same, a low complexity multiple dictionary based approach is proposed. This scheme exploits sparse coding to reduce the number of parameters to be estimated.

The rest of the paper is organized as follows: The proposed re-
dundant dictionary based adaptation approach is discussed in Section 2. The cluster-specific EV based adaptation technique is outlined in Section 3. The proposed multiple dictionary based approach is presented in Section 4. In Section 5, the experimental setup and evaluation of the proposed schemes is presented. The paper is finally concluded in Section 6.

2. Redundant dictionary based approach

2.1. Motivation

In the work reported in [15], a set of atoms is selected from a redundant dictionary using SR techniques. The selected bases are then linearly interpolated to derive the adapted model mean parameters for each test speaker/utterance. If $A_K$ denotes the matrix of the selected bases from the redundant dictionary $A$, where $\{a_i\}_{i=1}^K$ are the selected atoms, the Gaussian mean vector for a test speaker/utterance is then modeled as

$$\mu = \mu_0 + \sum_{i=1}^K w_i a_i$$  \hspace{1cm} (1)

where $\mu_0$ represents the SI mean supervector and $\{w_i\}_{i=1}^K$ is the set of basis interpolation weights. These interpolation weights are estimated using maximum likelihood eigen-decomposition (MLED) [8]. Except the Gaussian means, all the remaining parameters of the adapted model are borrowed from the SI model.

As discussed earlier, though the use of MP based SR is fast, it suffers from a major drawback that an atom may be selected more than once. This makes the equations involved in MLED unsolvable since the matrix $A_K$ becomes singular. In order to overcome this, the algorithm reported in [15] incorporates ways and means to detect and remove those redundant atoms. This incurs extra computations adding to the latency in the system response. Motivated by this, we explored bases selection using the orthogonal matching pursuit (OMP) algorithm [18]. In case of OMP, an atom can be selected only once and thus those extra computations are avoided.

2.2. OMP based basis search

To create a redundant dictionary, we first derive an SA model for each of the training speakers using mean-only MAP adaptation of the SI model. From each of these SA models, a corresponding supervector representation is extracted by the concatenation of the Gaussian mean parameters. Next, the eigenvoices are derived by performing PCA on the correlation matrix derived from these supervectors [19]. All these supervectors and the eigenvoices are collected in the form of a matrix $A$ referred to as a redundant dictionary which is used for the bases selection. For each test utterance, a set of atoms/bases is then dynamically selected from this dictionary using the OMP algorithm. To do so, the test utterance is also represented as Gaussian mean supervector (target) as explained in detail in [20]. OMP is then used to sparse code the target using the dictionary and those atoms are selected that have non-zero sparse coefficients. So selected atoms/bases are then interpolated to derive the adapted model mean parameters as done in earlier reported works [8, 15].

3. Cluster-specific Eigenvoices

As already discussed, the eigenvoices capture only the average acoustic/linguistic variations of the training data. This drawback can be mitigated by an acoustic clustering of the speaker space into subspaces. Unlike the conventional EV, the eigenvectors derived for a subspace will capture the local variations present in that cluster instead of the global variations. Interpolation of these cluster-specific eigenvectors to derive the adapted model will, in turn, lead to an improved performance. In this work, acoustic clustering is performed in the supervector domain using the vector quantization (VQ) technique. To do the same, the supervector representations are created for the training speakers as explained in Section 2.2. These supervectors are then grouped into $K$ clusters using VQ. For each of the $K$ sets of supervectors, a correlation matrix is created and PCA is performed on each of those matrices. $M$ eigenvectors corresponding to the largest eigenvalues are then selected from each of the sets. In the on-line adaptation phase, for each test utterance a set of $KM$ interpolation weights are jointly estimated using MLED. This approach is found to outperform the eigenvoices technique as discussed later in Section 5.

4. Multiple dictionary based approach

In the work reported in [15] and the approach proposed in Section 2, the SR techniques are used to select the atoms only. Model interpolation is done using weights derived via MLED. Motivated by this, we explored the means to directly use the sparse coded target as the Gaussian mean parameter in the adapted model. This will save the computation required for weight estimation which increases with an increase in the number of weights being estimated. In the following we first describe our attempts to use the sparse coded supervector as the adapted model mean parameters. This is followed by the description of the proposed low complexity model adaptation approach involving sparse coding over multiple dictionaries.

4.1. Using sparse coded supervector as model parameter

A comparative study of the nature of the sparse coefficients and the estimated interpolation weights reveals that they exhibit similar nature. In other words, for almost all the test utterances, when 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 magnitude of the sparse coefficients and the ML weights, almost same ordering is observed in both the cases. But there is a difference of a scaling factor between them, i.e., the sparse coefficients have a much lower dynamic range due to the normalization of atoms during OMP based search. Consequently, each dimension of the sparse coded target supervector also has a lower dynamic range than that of the corresponding dimension in the dictionary atoms. Such a supervector cannot be used as the Gaussian mean parameter in the adapted acoustic model. This is so because the dynamic ranges of the remaining model parameters (covariance, mixture-weights, etc.) are different as they had been jointly estimated with the unadapted mean parameter during training. In order to overcome this, we explored the derivation of a scaling factor so that the scaled sparse coded target can be used as the mean parameter in the adapted acoustic model. This can be easily done by a single point weight estimation using MLED.

The use of thus scaled supervector helped us to get only a slightly better performance over the unadapted SI system. In order to understand this behavior, we scrutinized the effect of initialization in the iterative MLED approach. For a robust estimation of weight parameters, unity weight is assigned to $a_0$ and zero to the rest as the initial values. In the first iteration, the posterior of the adaptation data with respect to the $w_i$th Gaussian component $\gamma_m(\tau)$ is computed using the SI model $a_0$. Using the current value of $\gamma_m(\tau)$, the weights are so updated that
the likelihood of the adaptation data increases. On successive iterations, a model is synthesized using the current value for the weights and \( \gamma_m(\tau) \) is re-computed from that model. The reason for the poor performance is that the posteriors do not change much on the second iteration of ML estimation as the sparse coded supervector has a much smaller dynamic range and hence the synthesized model is hardly any different from the SI model. The change in the likelihood of the adaptation data in successive iterations of the estimation process did confirm the same. In order to have better initial supervector, the obtained sparse coefficients are normalized so that they sum to one. The so derived sparse coded supervector, using the normalized sparse coefficients, has a similar dynamic range as that of the atoms. These modifications lead to a better initialization and hence a better performance. When this approach was used for the redundant dictionary case, an absolute improvement of 0.35% over the unadapted SI system was obtained.

4.2. Proposed approach

In above approach, the dynamic ranges of the eigenvoices and the speaker supervectors differ considerably. Hence, there is no precise control over the inclusion of intra- and inter-speaker information during the sparse coding of the target using a joint dictionary. To address this, we explored separate sparse coding over the eigenvoices dictionary and the dictionary composed of speaker supervectors and derived the corresponding sparse coded targets. This is followed by the joint scaling of the two sparse coded targets using MLED. In this way, the number of parameters can be considerably reduced while still capturing much finer acoustic space.

In order to generalize this approach, the number of dictionaries can be increased to obtain a greater number of sparse coded targets. To do the same, the earlier introduced cluster-specific eigenvoices can be used. Increasing the number of dictionaries implies an increase in the degree of freedom which, in turn, will result in improved performance. At the same time, with the sparse coding of the target, we are still able to contain the number of parameters required to be estimated and thus making it suitable for on-line applications. The proposed approach is outlined in Algorithm 1.

5. Results and Discussions

5.1. Experimental setup

To evaluate the performance of the proposed techniques on an LVCSR task, an ASR system is developed using HTK [21] on WSJCAM0 corpus [22]. In this corpus, the training set consists of 7861 utterances from 92 (male/female) speakers with approximately 90 sentences per speaker which corresponds to 15.5 hours of speech data. The test set consists 0.6 hours of speech data comprised of 308 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 state-clustered triphone is modeled using 8 Gaussians mixtures per state while 16 GMM/state is used for modeling the silence. A 5k-bigram language model is used in decoding. The baseline performance of the developed ASR system (the unadapted SI system) is 11.3%, same as that reported in [22]. The OMP-Box v-10 [23] is used for implementing OMP algorithm in Matlab. To simulate on-line adaptation task, adaptation is done for each of the test utterance individually. In other words, for each test utterance, a set of bases is selected and then interpolated independent of all other test utterance. Even the interpolation weights are estimated using the current test utterance only.

5.2. Results

The performances for the proposed redundant dictionary and cluster-specific EV based approaches along with that of the EV [8] and the improved reference speaker weighting (Im-RSW) [10], with variations in the number of the bases being interpolated are given in Figure 1. It is to note, that the proposed approaches outperform both EV and Im-RSW. The differences are not very large since utterance-specific unsupervised adaptation is performed with only 3-5 seconds of data. Similar relative improvements are also reported in [15] for the utterance-specific adaptation task performed using the WSJ database. In case of cluster-specific EV, the performances are shown for the
case when $K = 2$. In this case, the number of the cluster-specific speaker supervectors turn out to be 53 and 39. The performances of this approach, with variations in the number of covariance values derived for the clusters $M$ are shown in Figure 1. We tried to further extend the clustering of the speaker space into four subspaces. In case of WSJICAM0 data set, there are only 92 speakers in the training set, hence splitting them into four resulted in clusters with cardinality being 25, 20, 19 and 28 which is too less to provide an optimal learning for the cluster-specific EVs. As a result of this, we did not find any additional improvement with four clusters compared to that of the two cluster case. In future, we wish to try this extension on a speech database having a large number training speakers.

The redundant dictionary as well as the cluster-specific EV based approaches lead to the same best case WER of 10.45% (the best case WERs for EV and Im-RSW are 10.65% and 10.57%, respectively). The redundant dictionary based approach also requires a dynamic selection of bases which adds to the overall complexity. Both these approaches, though quite effective, still require the estimation of a large number of weight parameters (16 in the best case). As already discussed, this can be reduced by using the proposed multiple dictionary based approach. The performance evaluation results for the multiple dictionary based approach, for the cases of 2- and 4-dictionaries are shown in Figure 2. In case of 2-dictionary, the 92 speaker supervectors and their corresponding 92 eigenvectors are used, for the 4-dictionary case, the eigenvectors of the two cluster-specific speaker supervectors are considered. It is interesting to note that consistent improvements are obtained by increasing degree of freedom from 2 to 4. The best case WER for the 2- and 4-dictionary cases turn out to be 10.63% and 10.55%, respectively when sparsity is 12.

5.3. Discussion on complexity reduction

The proposed multiple dictionary based technique greatly reduces the computations involved during weight estimation since only two/four scaling factors are to be estimated. As the number of parameters to be estimated increases, the latency in the system response also increases due to the involved computations. For the best case scenario, in case of redundant dictionary based approach, 16 bases are selected from a dictionary of 184 atoms (92 SA and 92 EV) and hence the number of correlation computations performed is 2044 (= $16 \times 184$). On the other hand, for the 2-dictionary based approach, the number of correlation computations performed in sparse coding of the target is 2208 (= $12 \times 92 + 12 \times 92$) when the sparsity is set to 12. Similarly for the 4-dictionary case, this turns out to be 2208 (= $12 \times 53 + 12 \times 39 + 12 \times 53 + 12 \times 39$). Moreover, one can perform the multiple sparse codings in parallel once the target supervector is created. Consequently, the number of correlation computation will now reduce to 636 (= $12 \times 53$) for the 4-dictionary case. Table 1 presents a detail analysis of the complexities involved in the proposed techniques along with that for the EV based approach. In the best case scenario, the proposed multiple dictionary based approach results in a performance better than the existing techniques with the number of parameters estimated being reduced by a factor of 4.

6. Conclusion

A redundant dictionary based on-line adaptation approach has been presented in this paper. The discussed approach employs OMP algorithm for the selection bases from a redundant dictionary. This approach is found to give a relative improvement of 7.5% over the SI system when evaluated on an LVCSR task. Furthermore, we have also presented a low complexity fast adaptation approach involving sparse coding over multiple dictionaries. Consequently, only a few scaling factors are required to be estimated in place of a large number of interpolation weight parameters. For the best case scenario, the number of parameters being estimated is reduced by a factor of 4. Such a reduction is highly desirable especially in case of on-line tasks where the computational complexity is a major factor. Even with such considerable reductions in computational complexity, a relative improvement of 6.7% over SI is obtained.
7. References


