World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:6, 2007

Continuous Feature Adaptation for Non-Native Speech Recognition

Y. Deng, X. Li, C. Kwan, B. Raj, and R. Stern

***Abstract*—**The current speech interfaces in many military applications may be adequate for native speakers. However, the recognition rate drops quite a lot for non-native speakers (people with foreign accents). This is mainly because the non-native speakers have large temporal and intra-phoneme variations when they pronounce the same words. This problem is also complicated by the presence of large environmental noise such as tank noise, helicopter noise, etc. In this paper, we proposed a novel continuous acoustic feature adaptation algorithm for on-line accent and environmental adaptation. Implemented by incremental singular value decomposition (SVD), the algorithm captures local acoustic variation and runs in real-time. This feature-based adaptation method is then integrated with conventional model-based maximum likelihood linear regression (MLLR) algorithm. Extensive experiments have been performed on the NATO non-native speech corpus with baseline acoustic model trained on native American English. The proposed feature-based adaptation algorithm improved the average recognition accuracy by 15%, while the MLLR model based adaptation achieved 11% improvement. The corresponding word error rate (WER) reduction was 25.8% and 2.73%, as compared to that without adaptation. The combined adaptation achieved overall recognition accuracy improvement of 29.5%, and WER reduction of 31.8%, as compared to that without adaptation.

***Keywords*—**speaker adaptation; environment adaptation; robust speech recognition; SVD; non-native speech recognition

I. Introduction

EXISTING speech recognition software such as IBM via Voice or Dragon Naturally Speaking works well for native speaker and well-controlled background noise condition. However, the recognition performance drops quite significantly for non-native speakers and un-seen noisy conditions. The main reason is that the feature vectors of the non-native speech and corrupted speech are no longer similar to the distributions learned from the training data. This mismatch results in mis-classification and poor recognition [1], [2]. The training on accents and environmental condition specific data is impractical due the overhead of collecting large amount of data and unpredictable real environment.

To reduce the effect of mismatch, various techniques have been proposed in the literature, which can be broadly categorized as:

- Noise estimation and filtering that reconditions the speech signal or reconstruct speech features based on noise characteristics [3,4].
- On-line model adaptation to reduce the effect of mismatch in training and testing environments [5].
- Extraction of speech features robust to noise [6, 7], including features based on human auditory and perception modeling [8-10].
- Normalization techniques to compensate for the channel effect and speech production variations including cepstral mean normalization [11] and vocal track length normalization [12].
- Adaptation of acoustic model parameters to a specific speaker based on some criteria, including MLLR [13], constrained MLLR [14], maximum a posteriori (MAP) speaker adaptation [15], and Discriminative Likelihood Linear Transform (DLTR) [16].

Adaptation to the acoustic conditions of the test data is often performed by transforming the parameters of the state output densities of the Hidden Markov Models (HMM) in the recognizer through affine transforms [17]. While this method has proven to be highly effective, it is essentially an offline process that requires either significant amounts of adaptation data that are similar to the test data, or multiple passes over the test data. It also incurs the additional expense of having to transform the model parameters within the speech recognizer itself. Additionally, and possibly more importantly, the transformation parameters are learned such that the transformed model space best represents the complete set of incoming data from which they were learned. This does not account for any local variations that the data might have undergone even within the course of these recordings.

In this paper, we propose a novel continuous feature adaptation algorithm to compensate for training and testing data mismatch. The algorithm is general and can be applied to both noisy and non-native speech recognition. The salient features of this algorithm are:

1. Entirely based on acoustic feature space "tracking" and fully unsupervised.
2. Does not need the speaker-id information for adaptation, amenable to multi-users system.
3. Implemented by incremental singular value decomposition (SVD) [18] and runs in real time.

Manuscript received May 2, 2006. This work was supported by the U.S. Air Force under the Grant AF8650-05-C-6533

Y. Deng, X. Li, and C. Kwan, are with Intelligent Automation, Inc., Rockville, MD 20855 USA (Y. Deng is the corresponding author. phone: 301-294-5200; fax: 301-294-5201; e-mail: yunbin@jhu.edu).

B. Raj and R. Stern are with Carnegie Mellon University, Pittsburg, PA 15213 USA.

4. Continuously update transformation matrix based on a windowed acoustic features, effectively perform a non-linear transform and capable of capture local acoustic variations.
5. Can be further combined with other affine transform and normalization techniques.

The paper is organized as follows. Section II describes our novel continuous feature adaptation algorithm based on incremental SVD for automatic acoustic space "tracking". Section III reviews the well-known MLLR model adaptation algorithm and shows how to integrate our feature adaptation algorithm with the model adaptation algorithm. In Section IV, we will summarize the experimental results and the comparative studies. Finally, conclusions will be drawn in Section V.

II. Continuous Feature adaptation Algorithm

The mainstream acoustic features used for speech recognition is Mel Frequency Cepstral Coefficients (MFCC), which are obtained by taking the Discrete Fourier Transform (DCT) of log spectrum. The purpose of DCT is to achieve feature dimension reduction and de-correlation. It is well known that SVD projection results in the *most informative* subspace of all possible projection. While DCT uses constant transformation coefficient, the SVD needs to be trained from data and thus is sensitive to the training data acoustic characteristics. In this section, we first review the SVD technique and discuss the issues of applying it to speech recognition under mismatch condition; then we propose a continuous feature adaptation algorithm and its implementation by incremental SVD algorithm.

A. SVD for Dimension Reduction and De-correlation

The SVD is a technique for reducing the dimensionality of high-dimensional data sets [19]. Given an $d \times n$ data matrix M of rank r (where we assume, without loss of generality, that $d > n$), the SVD decomposes is given by

$$M_{d\times n} \to U_{d\times r} \cdot diag(s_{r\times 1}) \cdot (V_{r\times n})^T, \quad r \le \min(d,n), \tag{1}$$

where U and V are unitary matrix (i.e. $U^T = U^{-1}, V^T = V^{-1}$.), $diag(s)$ is an $r \times r$ diagonal matrix.

The columns of U represent the "eigenvectors" of M and represent a set of r orthogonal bases, and diagonal entries of $diag(s)$, termed the "singular values" of M, represent the scatter of the projections of the columns of M along the direction of these bases. SVD is often used to reduce the dimensionality of high-dimensional matrices. For instance, M may be reduced to a $k \times n$ matrix M' by projecting the columns of M the K columns of U that correspond to the K highest singular values in **S**

$$\tilde{M}_{k\times n} = (U_K)^T M_{d\times n}, \quad k < d, \tag{2}$$

where $U_K = U_{d\times k}$ is a matrix constructed from the K columns of U that correspond to the K highest singular values. One of the features of SVD is that a projection of the columns of M down to K dimensions in this manner is guaranteed to result in the most informative of all possible K dimensional projections of M.

B. SVD for Speech Recognition

Dimensionality reduction by SVD is frequently used in speech recognition systems to de-correlate and project relatively high-dimensional log-spectral vectors down to lower-dimensional cepstrum like feature vectors [20]. In order to do so, a large number of log-spectral vectors of a training data set are arranged in a matrix M, and the K-dimensional projection matrix U_K is derived by singular value decomposition of M. Thereafter U_K is used to project all log-spectral vectors, both from the training and test data, down to K-dimensional "cepstra", where, typically, K=13.

A problem arises when test data to be recognized are recorded in a different acoustic environment than the training data. In this case, the unitary projection matrix U_K is no longer guaranteed to be the most informative projection, resulting in a loss of crucial information in the test data, with subsequently lowered recognition performance. Independent projections cannot be derived for the test data since these projections may not conform to the original projections of the training data – in the worst case the independently learned projections from the test data might project them into an entirely different K-dimensional subspace from the training data. It therefore becomes necessary to identify a new projection matrix U'_K that de-correlates the test data *jointly* with the training data along the most informative directions.

C. Continuous Feature Adaptation Algorithm

We have developed a feature adaptation algorithm to continuously modify the incoming features to conform to the expected distribution of the training data. It composed of three steps: 1) train a transformation matrix from training data using principle component analysis (PCA); 2) adapt the transformation matrix by including current testing data using incremental SVD; 3) transform current testing data using this new transformation matrix. The continuous feature adaptation algorithm can be formulated as follows

Initialization

Initial transformation matrix, $T_{-1} = A$, is computed from Principal Component Analysis (PCA) of training data, i.e., the training feature matrix in log-spectral domain.

For t=0:T (on testing feature vectors)

$$T_t = Update(Downdate(T_{t-1}, X_{t-\tau}), X_t) \tag{3}$$

$$H_t = \gamma A + (1-\gamma) T_t \tag{4}$$

$$Y_t = H_t X_t, \tag{5}$$

where

X_t is the original testing feature vectors;

Y_t is the transformed testing feature vectors to be fed to the speech recognizer;

$\hat{T} = Update(T, X)$, is the function to update the transformation matrix *T* to include a new feature vector *X;*

$\hat{T} = Downdate(T, X)$ is the function to update the transformation matrix *T* to delete a history vector *X;*

τ is the "memory window" within which data vectors contribute to the current transformation;

$0 \le \gamma \le 1$ is a contribution tradeoff between training and current testing data;

H_t is the effective adaptation matrix used at time *t*.

As with other transformation-based algorithms, the algorithm is not specific to a particular type of acoustic condition and is equally effective in adapting to variations, both local and global, in noisy conditions, and speaker and accent variations. Since the algorithm is performed directly on incoming data, the transcriptions of the data are not required. That is, it is completely unsupervised. Furthermore, since transformation matrices are computed in an incremental and causal manner, the algorithm is well suited to run-time implementations.

Unlike current transformation-based data normalization algorithms [17], the transformation that is applied to each incoming vector is unique, since it is estimated causally from the entire sequence of incoming data vectors up to and including the current vector, but not including any vectors further downstream. The effect of such a transformation is twofold: 1) it projects the incoming test data into the same region of the data space that the training data are expected to lie, thereby increasing the probability of correct classification; 2) by normalizing the test data, it facilitates better estimation of model transformations for adaptation as the transformations need no longer account for data spread over a large region of the data space, resulting in improved recognition with transformed models. Since each vector is transformed uniquely, the effect of the data transformation is effectively non-linear and is not equivalent to that obtained with a single global affine transformation.

Direct implementation of such projection-learning mechanisms is, however, infeasible since it would require that the entire training data (or at least, sufficient statistics from it), can be retained and manipulated in conjunction with the test data in order to determine the new projections. In our work we circumvented this problem by adopting the incremental SVD algorithm proposed by Brand [1].

D. Implementation by Incremental SVD

The incremental SVD problem can be briefly stated as follows: given the SVD decomposition *U, S*, and *V* of a *dxn* matrix *M* and a new *dxc* matrix *C*, the goal is to obtain a new SVD matrix U'' that jointly de-correlates matrix **[***M C***]** without requiring explicit storage and manipulation of the original data matrix *M*. The incremental SVD algorithm proposed by Brand may be summarized as follows [18]:

The SVD of the training data is given by

$$M_{d\times n} \to U_{d\times r} \cdot diag(s_{r\times 1}) \cdot (V_{r\times n})^T, \quad r \le \min(d, n). \tag{6}$$

Given new testing samples $C_{d\times r}$, the matrix **[***M C***]** can be decomposed as follows

$$\begin{bmatrix} U & J \end{bmatrix} \begin{bmatrix} diag(s) & L \\ 0 & K \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & I \end{bmatrix}^T = [M \;\; C]. \tag{7}$$

where *J* is the orthogonal basis of *H,* for example, *J,K* could be a Q-R decomposition of *H*, specifically

$$\begin{aligned} L &= U^T C \\ H &= C - UL \\ K &= J^T H. \end{aligned} \tag{8}$$

The middle matrix *Q* is diagonal with a c-column border, which needs to be further diagonalized. This is done using SVD again. Since *Q* is a small matrix, this SVD can be done very efficiently.

$$Q = \begin{bmatrix} diag(s) & L \\ 0 & K \end{bmatrix} \to U' \cdot diag(s') \cdot (V')^T \tag{9}$$

The final decomposition matrices are given by

$$\begin{aligned} U'' &= [U \;\; J] \cdot U' \\ s'' &= s' \\ V'' &= \begin{bmatrix} V & 0 \\ 0 & I \end{bmatrix} V' \end{aligned} \tag{10}$$

It is easy to verify that

$$\begin{aligned} U'' \cdot diag(s'') \cdot (V'')^T &= [U \cdot diag(s) \cdot (V)^T \;\; C] \\ &= [M \;\; C]. \end{aligned} \tag{11}$$

A special case is when the additional data matrix *C* is a single vector $c = C$. The computation can be done very quickly since *K* becomes a scalar, $k = K = \left\| c - UU^T c \right\|$, and *J* becomes a vector, $j = J = (c - UU^T c)/k$. This is what we implemented in our feature adaptation algorithm.

Note that the above derivation only requires U_K, *J, S* and *C* to compute *U"* and the training data *M* are not required at all. Only the first *k* columns of *U"* are retained for the projection.

The incremental SVD algorithm described above can now be utilized to compute the optimal projection for a test data *C* as follows: Compute an initial projecting eigen-matrix U_K by SVD of the training data. On the test data, we compute a new projection matrix $U"_K$ by incrementally considering *C* in conjunction with the test data as described above. In practice, *C* may be up or down weighted in the data matrices, prior to update, via a recursive IIR filter formulation that is permitted within the incremental SVD algorithm. Then we use the new

matrix $U"_K$ to project C down for recognition. Note that the new projection matrix $U"_K$ includes residual information from subspaces not included in the training data; hence additional adaptation of the models using MLLR may be expected to be beneficial for this procedure.

The incremental SVD algorithm incrementally computes the optimal orthogonalizing transformation for a data set. The algorithm has the following important features: 1) it incrementally updates the transformation matrix with each additional vector incorporated into the data set, without requiring explicit storage of the entire data set, 2) it incrementally updates the transformation matrix to account for deletion of vectors from the data set, also without requiring explicit storage of the entire data set, and 3) it performs both operations in linear time, as opposed to the quadratic time required by most SVD algorithms.

III. An Integraed Approach to Imporve Speech Recognition Accuracy for Non-native Spakers

The feature adaptation method is different from the traditional Weiner filter and KLT (principal components analysis) for feature space transformations. Weiner Filtering works when the noise estimate is reliable regardless of how well the speech matches the clean training speech. KLT works when the speech components of the noisy data match the clean speech training data. In the current scenario, neither a reliable noise estimate is available, nor is a clean speech prior available. When a speaker changes, the accent may change. Hence, the acoustics of each distinct unit may change, and there may or may not be significant ambient noise in the background. The continuous feature adaptation algorithm presented above is observed to be able to improve recognition in all these situations.

Salient features of this technology are:

- The features are entirely based on acoustic space "tracking", starting from a set of initial transformation vector values which are predetermined for a group of *accents* in the our study (the initial vectors would be predetermined for a set of *speakers* in a speaker verification task, for example, or on a set of *noisy conditions*, in a noise attenuation task – a combination of such conditions is also possible).
- The tracking is based on the principle of incremental linear time SVD (singular value decomposition) described by Brand [18]. It runs real time on all current processors.
- Continuously update transformation matrix based on a windowed acoustic features, effectively perform a non-linear transform and capable of capture local acoustic variations.
- Can be further combined with other affine transform and normalization techniques.

Among many speaker adaptation algorithms, the Maximum Likelihood Linear Regression (MLLR) is most widely used and has shown to significantly improve speech recognition accuracy for accented speech using very few adaptation data [21-23]. Even though many advanced model adaptation algorithms have been developed recently [29-31], as a proof of concept we choose the basic MLLR algorithm for its simplicity. In this section, we first briefly review the MLLR algorithm and then propose an integrated feature and model adaptation system for non-native speech recognition.

A. MLLR Model Adaptation Algorithm

Due to limited amount of adaptation data, the MLLR algorithm usually only updates Gaussian mean vectors by a linear transformation. The distribution of data is assumed to be K mixtures of Gaussian. If a baseline mean vector is u_k, the corresponding adapted mean vector $\hat{u}_k$ for a new speaker or environmental condition is given by linear regression parameters A and b,

$$\hat{u}_k = Au_k + b, \qquad 1 \le k \le K \tag{12}$$

The task is to estimate A and b such that the likelihood of adaptation data $O = (o_1, o_2, ..., o_T)$ is maximized. Assume u_k, $\hat{u}_k$ and b are D dimension vector. It follows

$$N(\hat{u}_k, C_k) = N(Au_k + b, C_k), \quad 1 \le k \le K, \tag{13}$$

$$C_k = diag(\sigma_{k1}^{-2}, \sigma_{k2}^{-2}, ..., \sigma_{kD}^{-2}) \tag{14}$$

It has been shown [13] that we can estimate A and b by minimizing

$$Q = \sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_t(k)(o_t - Au_k - b)^T C_k^{-1}(o_t - Au_k - b) \tag{15}$$

where $\gamma_t(k) = \gamma_t(s,k)$ is the posterior probability of being in state s at time t with k^{th} Gaussian. A and b can then be solved by setting the derivatives of Q with respective to A and b to zero.

B. Combine Feature and MLLR Model Adaptation

As described above, both model-based MLLR and feature based continuous adaptation can improve speech recognition rate significantly. Since these two methods are independent, we can combine them to improve the speech recognizer further. We first perform feature-based continuous adaptation, and then implement model-based MLLR on the new speech feature. Fig. 1 shows the integrated approach. Basically, the feature based method improves the Mel Frequency Cepstral Coefficients (MFCC). Then, the improved feature vector goes into the MLLR algorithm for updating the speaker specific acoustic model parameters. Finally, the new parameters go into the speech engine. Note, both the feature adaptation and MLLR model adaptations algorithms we implemented are unsupervised. For the MLLR algorithm, only a global transformation matrix for the Gaussian mean vectors is trained.

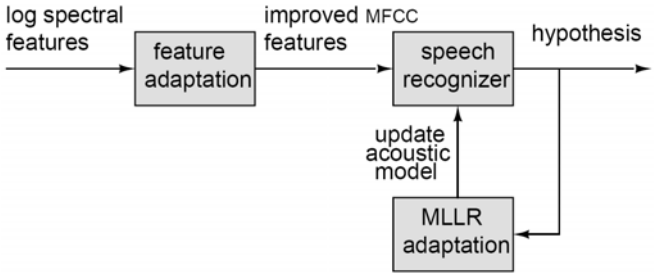

Fig. 1 Block diagram of the combined approach for speaker adaptation

IV. Experimental results

In this section we describe the evaluation of the feature and model adaptation algorithm on NATO non-native speech data base.

A. The NATO Non-native Speech Database

The NATO native and non-native corpus was developed by a NATO research group to provide a military oriented database for multilingual and non-native speech processing studies [24]. Speech data was recorded in naval transmission training center of four countries Germany (GE), Netherlands (NL), United Kingdom (UK), and Canada (CA). The subjects from Germany, Netherlands and UK were native speakers of German, Dutch and UK English, respectively. The Canadian subjects included native speakers of both English and Canadian French. Every speaker recorded a number of utterances in the international argot of the air force (English), as well as a rendition of Aesop's fable "The Northwind and the Sun", in both their native language and English. In this paper, recognition was performed only on the English utterances in the database. The detailed data and speaker information is given in TABLE .

TABLE I
NATO DATA AND SPEAKER INFORMATION

	CA	GE	NL	UK
Data (hours)	2.49	2.25	2.53	1.63
# Speakers	22	51	48	13
# Women	5	0	9	5
Age	22-35	17-23	17-61	19-62
Average data per speaker (minutes)	6.79	2.65	3.16	7.52

B. The baseline Speech Recognition System

1) Acoustic Model (AM) Training

The Carnegie Mellon University (CMU) Sphinx-3 continuous density HMM system was used for our study [25]. HMMs with 5000 tied states, each modeled by a mixture of 8 Gaussians, were trained from native American speech: 130 hours of BN (broadcast news) data combined with 33 hours of SPINE1 (Speech in Noisy Environments 1) and SPINE2 data [26] [27].

The CMUdict pronunciation dictionary was used for the experiment. The pronunciations in this dictionary represent standard American pronunciation of all words, expressed in terms of 40 phonemes. No lexical adaptation was done.

2) Language Model Training

In order to bring out the effect of acoustic model adaptation algorithm, we first used simple uni-gram language model (LM) for speech recognition. The uni-gram model accorded equal probability to all the words in the recognition vocabulary. A tri-gram LM is trained with probability masses redistributed by the Good Turing discounting strategy [28]. We randomly partitioned the NATO data into two parts, part A and part B, which are roughly equal in terms of data size. When we performed recognition of part A, a language model trained from part B was adopted, and vice versa.

C. Speech Recognition Experimental Results

1) Baseline Results

Using the above acoustic model trained from native data and NATO N4 non-native speech corpus as a test set [1], the baseline performance is summarized in TABLE . Due to the well-defined language structure in military communication, the tri-gram language model outperformed uni-gram language model significantly.

TABLE II
BASELINE NATO SPEECH RECOGNITION ACCURACY WITH UNIGRAM AND TRIGRAM LANGUAGE MODELS

Data	Baseline with unigram (%)	Baseline with trigram (%)	Relative improvement (%)
CA	57.26	77.75	35.78
GE	26.85	52.51	95.57
NL	34.22	59.90	75.04
UK	46.03	69.25	50.44

2) Continuous Feature Adaptation Results

In feature adaptation experiments, non-native NATO log spectral features are continuously transferred before passing to the SPHINX decoder. TABLE summarizes the results of the feature based adaptation method. Tri-gram language model was used in this experiment. In the table, the numbers outside the parentheses represent recognition accuracy, which corresponds to the commonly used metric of recall. The numbers within parentheses represent the recognition error, which is the sum of substitutions, deletions, and insertions error. The proposed feature based adaptation algorithm improved the baseline performance by an average of 15%. Performance on German and Dutch speakers has been improved the most.

TABLE III
PERFORMANCE OF THE PROPOSED CONTINUOUS FEATURE BASED ADAPTATION METHOD (WITH TRIGRAM LM)

Data	Baseline (%)	With Feature adaptation (%)	Relative improvement (%)
CA	77.8 (33.4)	82.6 (27.3)	6.2 (18.3)
GE	52.5 (55.8)	64.6 (42.3)	23.0 (24.2)
NL	59.9 (49.7)	73.0 (33.0)	21.9 (33.7)
UK	69.3 (49.6)	76.5 (36.2)	10.5 (26.9)

3) MLLR Model Adaptation Results

Our study shows the supervised approach offers about 1% of improvement over the unsupervised one on NATO data. Since speakers are usually reluctant to perform supervised training, our experiments were all based on unsupervised adaptation. Table 4 summarizes the experimental results. The MLLR adaptation algorithm improves the word recognition accuracy by an average of 11%. Similar to feature adaptation, performance on German and Dutch speakers has been improved the most. It might due to the fact that German and Dutch accent are more different from American accent. The WER reduction does not improve as much as the recognition accuracy. This is attributable to the fact that the models for the background (non-speech) were adapted with the same matrices as the models for speech. This results in the insertion of a large number of spurious words in the recognition hypothesis in non-speech segments, as well as the misrecognition of several of the uttered words as silence.

TABLE IV
WORD RECOGNITION ACCURACY OF BASELINE AND THE MLLR ADAPTATION ON NATO DATABASE (WITH TRIGRAM LM)

Data	Baseline (%)	With MLLR adaptation (%)	Relative improvement (%)
CA	77.8 (33.4)	83.9 (30.5)	7.9 (8.8)
GE	52.5 (55.8)	64.0 (60.0)	11.5 (-7.1)
NL	59.9 (49.7)	71.0 (47.0)	18.6 (5.5)
UK	69.3 (49.6)	74.9 (47.7)	8.1 (3.7)

4) Combined Adaptation Results

TABLE and Table 6 summarize the performance by combining the feature and MLLR model adaptation algorithms, with the unigram and trigram LM, respectively. All adaptation algorithms are unsupervised. With the tri-gram LM, The average overall accuracy improvement is 29.5% and the WER reduction is 31.8%.

TABLE V
RECOGNITION ACCURACY (WITH UNIGRAM LM)

Data	Baseline (%)	Feature adaptation (%)	MLLR adaptation (%)	Combined adaptation (%)
CA	57.3	63.5	65.7	71.0
GE	26.9	34.0	38.8	44.8
NL	34.2	45.3	47.8	55.8
UK	46.0	53.1	51.8	58.4

TABLE VI
RECOGNITION ACCURACY (WITH TRIGRAM LM)

Data	Baseline (%)	Combined adaptation (%)	Overall Improvement (%)
CA	77.8 (33.4)	86.9 (25.6)	11.7 (23.3)
GE	52.5 (55.8)	76.8 (40.8)	46.3 (27.0)
DL	59.9 (49.7)	83.3 (27.2)	39.1 (45.2)
UK	69.3 (49.6)	83.8 (33.9)	20.9 (31.5)

D. Comparative Studies

As a comparative study, we trained a custom tied-state tri-phone acoustic model for United Kingdom speaker from the WSJCAM0 database (CD0 and CD1 data). The WSJCAM0 is a British English speech corpus derived from Wall Street Journal text corpus [32]. Using the same language model, the recognition on UK part of NATO data shows a recognition accuracy of 78.9% and WER of 34.8%. The combined feature and model adaptation algorithm using acoustic model trained from native American speech achieves accuracy of 83.8% and WER of 33.9% (shown in Table 6). Our integrated adaptation approach for non-native speech recognition outperforms the custom accent specific acoustic model. The time-consuming data collection can be avoided by our integrated speaker adaptation algorithm.

V. CONCLUSIONS

In the paper, a feature based adaptation algorithm was proposed for unsupervised continuous speaker and environmental adaptation. Like MLLR, which modifies acoustic models to reduce the mismatch between training and test conditions, feature based adaptation also reduces the mismatch between the intra-phoneme spectral variations that occur as a result of non-nativity in the test data, as compared to those encountered in the training data. Experiments on NATO non-native database has shown significant speech recognition accuracy improvement over baseline acoustic model trained on native American English speaker. The feature based adaptation integrated with MLLR model based adaptation improved the performance even further.

The feature based adaptation algorithm described in Section III is not specific to a particular type of acoustic condition and is equally effective at adapting to variations, both local and global, in noisy conditions, speaker, and accent variations. However, it is critical to initialize the matrix appropriately (completely random initializations do not work). In our current work, the initial transformation matrix A was determined using data from a group of "typical" American data with accents, since our focus is on accent robustness. In a condition-specific task, it is more appropriate to initialize with the condition specific data. For example, in a task where robustness to speakers of a single accent is important, A could be estimated using data from many speakers, spanning, as far as possible, the range of variations expected in the vocal tract characteristics of test speakers. In a noise attenuation task, the initial transformation matrix Acould be estimated using representative noise conditions. In mixed-focus tasks, some appropriate combination of data conditions could be used.

ACKNOWLEDGMENT

The authors would like to thank Mr. David Williamson (the topic chairman for this program) of the Air Force for encouragements and financial support and Dr. Rita Singh of Haikya Corporation for many constructive discussions and support throughout this research.

REFERENCES

[1] B. R. Ramakrishnan, *Recognition of Incomplete Spectrograms for Robust Speech Recognition*, Ph.D. dissertation, Dept. Electrical and Computer Engineering, Carnegie Mellon University, 2000.

[2] Z. Wang, T. Schultz, A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech" *IEEE Int.. Conf. Acoust. Speech Signal Process (ICASSP)*, 2003.

[3] S.V., Milner, B.P, "Noise-adaptive hidden Markov models based on Wiener filters", *Proc. European Conf. Speech Technology*, Berlin, 1993, Vol. II, pp.1023-1026.

[4] "Acoustical and Environmental Robustness in Automatic Speech Recognition". A. Acero. Ph. D.Dissertation, ECE Department, CMU, Sept. 1990.

[5] Nadas, A., Nahamoo, D. and Picheny, M.A, "Speech recognition using noise-adaptive prototypes", *IEEE Trans. Acoust. Speech Signal Process.* Vol.37, No. 10, pp-1495- 1502, 1989.

[6] Mansour, D. and Juang, B.H, "The short-time modified coherence representation and its application for noisy speech recognition", *Proc. IEEE Int.. Conf. Acoust. Speech Signal Process.*, New York, April 1988.

[7] S. Chakrabartty, Y. Deng and G. Cauwenberghs, "Robust Speech Feature Extraction by Growth Transformation in Reproducing Kernel Hilbert Space," *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP'2004)*, Montreal Canada, May 17-21, 2004.

[8] Ghitza, O., "Auditory nerve representation as a basis for speech processing", in *Advances in Speech Signal Processing*, ed. by S. Furui and M.M.Sondhi (Marcel Dekker, New York), Chapter 15, pp.453-485.

[9] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoustic Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, Apr. 1990.

[10] Y. Deng, S. Chakrabartty, and G. Cauwenberghs, "Analog Auditory Perception Model for Robust Speech Recognition," *Proc. IEEE Int. Joint Conf. on Neural Network (IJCNN'2004)*, Budapest Hungary, July 2004.

[11] F.H. Liu, R.M. Stern, X. Huang, A. Acero, "Efficient Cepstral Normalization for Robust Speech Recognition", Proceedings of ARPA Speech and Natural Language Workshop, 1993.

[12] S. Wegmann, D. McAllaster, J. Orloff, B. Peskin, "Speaker normalization on conversational telephone speech", *Proc. ICASSP*, 1996.

[13] C. J. Leggetter, P. C. Woodland, "Speaker adaptation of HMMs using linear regression", *Technical Report CUED/F-INFENG/ TR. 181*, Cambridge University, 1994.

[14] D. Giuliani, M. Gerosa, F. Brugnara, "Speaker Normalization through Constrained MLLR Based Transforms", *International Conference on Spoken Language Processing, ICSLP*, 2004.

[15] C.H. Lee, J.L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters", *Acoustics, Speech, and Signal Processing, ICASSP*, 1993.

[16] V. Doumpiotis, S. Tsakalidis, and W. Byrne. "Discriminative linear transforms for feature normalization and speaker adaptation in HMM estimation", *IEEE Transactions on Speech and Audio Processing*, 13(3), May 2005.

[17] G. Saon, G. Zweig and M. Padmanabhan, "Linear feature space projections for speaker adaptation", ICASSP 2001, Salt Lake City, Utah, 2001.

[18] Brand, M., "Incremental singular value decomposition of uncertain data with missing values", *Proceedings, European Conference on Computer Vision*, ECCV, 2002.

[19] Ed. F. Deprettere, SVD and Signal Processing: Algorithms, Analysis and Applications, Elsevier Science Publishers, North Holland, 1988.

[20] K. Hermus, I. Dologlou, P. Wambacq and D. V. Compernolle**.** "Fully Adaptive SVD-Based Noise Removal for Robust Speech Recognition", In *Proc. European Conference on Speech Communication and Technology*, volume V, pages 1951--1954, Budapest, Hungary, September 1999.

[21] L. F. Uebel and P. C. Woodland, "Improvements in linear transforms based speaker adaptation," in *ICASSP*, 2001.

[22] T. Anastasakos, J. McDonough, R. Schwartz, etc, "A compact model for speaker-adaptive training," in *ICSLP*, 1996.

[23] P. C. Woodland and D. Povey, "Large scale discriminative training for speech recognition," in *Proceedings of the Tutorial and Research Workshop on Automatic Speech Recognition*. *ISCA*, 2000.

[24] L. Benarousse, E. Geoffrois, J. Grieco, R. Series, etc,, "The NATO Native and Non-Native (N4) Speech Corpus", in *Proceedings Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark, 2001.

[25] M. K. Ravishankar, "Sphinx-3 s3.3 Decoder", Sphinx Speech Group, CMU.

[26] P Beyerlein, X Aubert, R Haeb-Umbach, M Harris, "Large vocabulary continuous speech recognition of Broadcast News–The Philips/RWTH approach", *Speech Communication*, 2002.

[27] V.R. Gadde, A. Stolcke, D. Vergyri, J. Zheng, K. Sonmez, "Building an ASR System for Noisy Environments: SRI's 2001 SPINE Evaluation System", *Proceedings of ICSLP*, 2002.

[28] S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 35(3), pp. 400-401, March, 1987.

[29] D. Povey, P.C. Woodland, M.J.F. Gales, "Discriminative MAP for acoustic model adaptation", *Proc. ICASSP*, 2003.

[30] J. Stadermann and G. Rigoll, "Two-stage speaker adaptation of hybrid tied-posterior acoustic models," in *ICASSP*, 2005.

[31] P. Kenny, G. Boulianne, P. Dumouchel, "Eigenvoice Modeling with Sparse Training Data", *IEEE Transactions on Speech and Audio Processing*, 2005.

[32] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, "Wsjcam0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition ", *Proc. ICASSP*, 1995.

Yunbin Deng received his B.S. degree in Control Engineering from Beijing University of Aeronautics and Astronautics in 1997, his M.S. in Electrical Engineering from Institute of Automation, Chinese Academy of Sciences in 2000, and his M.S.E. in Electrical and Computer Engineering (ECE) form Johns Hopkins University (JHU) in 2002, respectively. He received his Ph.D. at ECE, JHU in April 2006. He joined Intelligent Automation, Inc. in Rockville, Maryland as a research engineer in 2005. Dr. Deng's research interests include language and speech processing, robust and non-native speech recognition, dialog system, mixed-signal VLSI circuits and system, and machine learning. He won Outstanding Overseas Chinese Student Award by Chinese Scholarship Council at 2003. He is a student member of the IEEE and SPIE. Mr. Deng is a manuscripts reviewer for IEEE Transaction on Audio and Speech Processing, IEEE Transaction on Circuit and System, and Circuit, System, and Signal Processing Journal.

Xiaokun Li received his BS and MS degree in electrical engineering from Xian Jiaotong University, China, 1992 and 1995 respectively. He obtained his Ph.D. degree in electrical engineering from University of Cincinnati, Ohio, in 2004. From 1995 to 1999, he was an assistant professor at Xian Jiaotong University. He worked as a visiting researcher at SCR (Siemens Corporate Research), Princeton, NJ, in 2002, and MERL (Mitsubishi Electronic Research Labs), Cambridge, MA, in 2003. Since 2004, he has worked with Intelligent Automation Inc as a research engineer. His research interests include signal/image processing and analysis, optical/electronic imaging, medical imaging, computer vision, machine learning, pattern recognition, artificial intelligence, real-time system, and data visualization.

Chiman Kwan received his B.S. degree in electronics with honors from the Chinese University of Hong Kong in 1988 and M.S. and Ph.D. degrees in electrical engineering from the University of Texas at Arlington in 1989 and 1993, respectively. From April 1991 to February 1994, he worked in the Beam Instrumentation Department of the SSC (Superconducting Super Collider Laboratory) in Dallas, Texas, where he was heavily involved in the modeling, simulation and design of modern digital controllers and signal processing algorithms for the beam control and synchronization system. He received an invention award for his work at SSC. Between March 1994 and June 1995, he joined the Automation and Robotics Research Institute in Fort Worth, where

he applied neural networks and fuzzy logic to the control of power systems, robots, and motors. Since July 1995, he has been with Intelligent Automation, Inc. in Rockville, Maryland. He has served as Principal Investigator/Program Manager for more than 65 different projects, with total funding exceeding 20 million dollars. Currently, he is the Vice President of IAI. He has published more than 40 papers in archival journals and has had 100 additional refereed conference papers. He is a senior member of the IEEE.

Bhiksha Raj received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, Pennsylvania, in May 2000. Since 2001, he has been working at Mitsubishi Electric Research Laboratories, Cambridge, Massachusetts. He works mainly on algorithmic aspects of speech recognition, with special emphasis on improving the robustness of speech recognition systems to environmental noise. His latest work was on the use of statistical information encoded by speech recognition systems for various signal processing tasks. He is a member of the IEEE.

Richard M. Stern received the S.B. degree from the Massachusetts Institute of Technology (1970), the M.S. degree from the University of California, Berkeley (1972), and the Ph.D. from MIT (1977), all in electrical engineering. He has been on the faculty of Carnegie Mellon University since 1977, where he is currently a professor in the Electrical and Computer Engineering, Computer Science, and Biomedical Engineering Departments and the Language Technologies Institute. Much of his current research is in spoken language systems. He is particularly concerned with the development of techniques to make automatic speech recognition more robust with respect to changes in environmental and acoustical ambience. He has also developed sentence parsing and speaker adaptation algorithms for earlier CMU speech systems. In addition to his work in speech recognition, he also maintains an active research program in psychoacoustics, where he is best known for theoretical work in binaural perception. He is a member of the IEEE.