AN INTEGRATED APPROACH TO IMPROVE SPEECH RECOGNITION RATE FOR NON-NATIVE SPEAKERS

Y. Deng, X. Li, C. Kwan, R. Xu
Intelligent Automation, Inc.
ydeng,xli,ckwan,hgxu@i-a-i.com

B. Raj, R.Stern
MERL,Carnegie Mellon University
bhiksha@merl.com, rms@cs.cmu.edu

D. Williamson
Wright-Patterson AFB
david.williamson@wpafb.af.mil

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 loud environmental noise such as tank noise, helicopter noise, etc. In this paper, we proposed a novel speech feature adaptation algorithm for continuous accent and environmental adaptation. 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 combined adaptation achieved overall recognition accuracy improvement of 29.5%, and word error rate reduction of 31.8%.

Index Terms: Non-native speech recognition, feature adaptation, model adaptation, accent and speaker adaptation

10.21437/Interspeech.2006-481

1 INTRODUCTION: SVD FOR SPEECH RECOGNITION

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. 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.

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 [1]

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

where U and V are unitary matrix, $diag(s)$ 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.

Dimensionality reduction by SVD is frequently used in speech recognition systems to de-correlate and project high-dimensional log-spectral vectors down to lower-dimensional cepstrum like feature vectors [2]. 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*. 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.

2 CONTINUOUS UNSUPERVISED FEATURE ADAPTATION

We 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 SVD of training data, i.e., the MFCC feature vectors.

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

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

$$H_t = \gamma A + (1-\gamma)T_t \quad (4)$$

$$Y_t = H_t X_t, \quad (5)$$

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.

It is critical to initialize the matrix A appropriately. In our current work, the initial value of A was determined using data from a group of typical Native American data with *accents,* since our focus is on accent robustness.

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, speaker and accent variations. Since the algorithm is performed directly on incoming acoustic data and needs no transcriptions, 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 [3], 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 other model adaptation technique, 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 a single global affine transformation.

Direct implementation of such SVD projection-learning mechanisms is, however, infeasible since it would require that the entire training data (or at least, sufficient statistics from it) 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 [4].

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 can be summarized as follows [4]:

The SVD of the training data is given by

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

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

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

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

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

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.

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

The final decomposition matrices are given by

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

It is easy to verify that

$$U'' \cdot diag(s'') \cdot (V'')^T = [U \cdot diag(s) \cdot (V)^T \;\; C] \quad (11)$$
$$= [M \;\; C].$$

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.

Salient features of this technology 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) and runs in real time.
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.

3 COMBINING MODEL-BASED MLLR WITH FEATURE-BASED APPROACH

Previous research on speaker adaptation has been focused on model adaptation and pronunciation adaptation. Since the continuous feature adaptation and the model adaptation algorithms are independent, we can be combined.

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 [5-7]. Even though many advanced model adaptation algorithms have been developed recently [8-10], as a proof of concept we choose the basic MLLR algorithm for its simplicity. The MLLR we implemented is an unsupervised single class transform applied only to the Gaussian means.

We first perform feature-based continuous adaptation, and then implement model-based MLLR on the new speech feature. Figure 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 model parameters. Finally, the new parameters go into the speech engine.

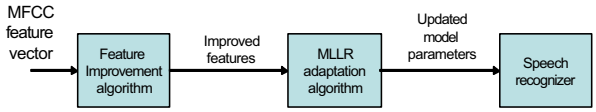

Figure 1 Block diagram of the combined approach for speaker adaptation.

4 EXPERIMENT ON NATO DATABASE

4.1 The NATO 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 [11]. 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, for a total 2223 utterances and 50.7K words. The vocabulary size is about 1K. The detail data and speaker information are given in Table 1.

	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

Table 1 NATO data and speaker information

4.2 Acoustic Model and Language Model Training

The Carnegie Mellon University (CMU) Sphinx-3 continuous density Hidden Markov Model (HMM) system was used for our study. HMMs with 5000 tied states, each modelled 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.

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.

A tri-gram LM is trained with probability masses redistributed by the Good Turing discounting strategy. 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.

4.3 Experimental results

Our following experiments are all based on unsupervised adaptation. In all tables, the numbers outside the parentheses represent recognition accuracy, while the numbers within parentheses represent the recognition error, all in percentage. Table 2 summarizes the experimental results. The MLLR adaptation algorithm improves the word accuracy (only consider deletions and substitution error) by an average of 11%. The WER (also consider the insertion error) 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.

Data	Baseline (%)	With MLLR adaptation	Relative improvement
CA	77.75 (33.39)	83.92 (30.46)	7.94 (8.78)
GE	52.51 (55.84)	63.97 (59.78)	11.46 (-7.06)
NL	59.90 (49.70)	71.02 (46.98)	18.56 (5.47)
UK	69.25 (49.55)	74.86 (47.70)	8.10 (3.73)

Table 2: Baseline and after MLLR results.

Table 3 summarizes the results of the feature adaptation method. The input to the feature adaptation is the standard 39-dimension MFCC feature, while the output is improved 39-dimentison MFCC feature. The proposed feature adaptation algorithm improved the baseline performance by an average of 15%. Similar to MLLR, performance on German and Dutch speakers has been improved the most. It might due to the fact that German and Dutch accent are quite different from American accent.

Data	Baseline (%)	With Feature adaptation	Relative improvement
CA	77.75 (33.39)	82.58 (27.28)	6.21 (18.3)
GE	52.51 (55.84)	64.56 (42.33)	22.95 (24.19)
NL	59.90 (49.70)	73.00 (32.96)	21.87 (33.68)
UK	69.25 (49.55)	76.50 (36.22)	10.47 (26.9)

Table 3: Baseline and after feature adaptation results.

Table 4 summarizes the performance by combining the feature and model adaptation algorithms. The average overall improvement is 29.5%.

Data	Baseline (%)	Combined adaptation	Overall Improvement
CA	77.75 (33.39)	86.85 (25.61)	11.7 (23.30)
GE	52.51 (55.84)	76.84 (40.75)	46.33 (27.02)
NL	59.90 (49.70)	83.30 (27.23)	39.07 (45.21)
UK	69.25 (49.55)	83.75 (33.92)	20.9 (31.54)

Table 4: Integrated system recognition accuracy.

5 CONCLUSIONS

In the paper, a feature based adaptation algorithm was proposed for unsupervised continuous speaker and environmental adaptation. Experiments on NATO non-native database has shown significant speech recognition accuracy improvement over baseline acoustic model trained on native English speaker. The feature based adaptation integrated with MLLR model based adaptation improved the performance even further.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Rita Singh at Haikay Corp. for many constructive discussions.

REFERENCES

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

[2] 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.

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

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

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

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

[7] 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.

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

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

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

[11] 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.