interspeech 2008 incorporating SST 08

Environment Mismatch Compensation using Average Eigenspace for Speech Recognition

Abhishek Kumar, John H. L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas-75083, USA
abhik@student.utdallas.edu, john.hansen@utdallas.edu

Abstract

The performance of speech recognition systems is adversely affected by mismatch in training and testing environmental conditions. In addition to test data from noisy environments, there are scenarios where the training data itself is noisy. Speech enhancement techniques which solely focus on finding a clean speech estimate from the noisy signal are not effective here. Model adaptation techniques may also be ineffective due to the dynamic nature of the environment. In this paper, we propose a method for mismatch compensation between training and testing environments using the "average eigenspace" approach when the mismatch is non-stationary. There is no need for explicit adaptation data as the method works on incoming test data to find the compensatory transform. This method is different from traditional signal-noise subspace filtering techniques where the dimensionality of the clean signal space is assumed to be less than the noise space and noise affects all dimensions to the same extent. We evaluate this approach on two corpora which are collected from real car environments: CU-Move and UTDrive. Using Sphinx, a relative reduction of 40-50% is achieved in WER compared to the baseline system. The method also results in a reduction in the dimensionality of the feature vectors allowing for a more compact set of acoustic models in the phoneme space.

Index Terms: speech recognition, feature adaptation, eigenvector, simultaneous diagonalization

1. Introduction

The performance of Automatic Speech Recognition (ASR) systems suffers dramatically when there is a mismatch in training and test data conditions. This mismatch can be due to many reasons including changes in background noise, changes in training and test recording conditions (different microphone, channel effects) which result in convolutive mismatch, changes in speaker accents (native/non-native speakers) and speaker stress levels, etc. Further, in real time car environments which require hands-free speech recognition, the test data conditions change at a relatively higher rate with changes in speed, accompanying traffic and car window positions (rolled up or down) and other factors, so the test environment is inherently non-stationary. This paper focuses on the problem of reducing the mismatch between train and test environments and data conditions where this mismatch is non-stationary.

This project was funded in part by NEDO through its research management agreement with CRSS-UTD, and in part by the Univ. of Texas at Dallas under project EMMITT.

Considerable research has been conducted on the problem of mismatch compensation for speech recognition. Some of these schemes require the presence of stereo data (simultaneous recording from both environments) which is not available in most real scenarios. Other approaches do not mandate the availability of stereo data, but they require some information about the environments such as a model or knowledge of environmental statistics. These approaches can be classified under two broad categories: model based methods and feature based methods. Model based approaches try to transform the phoneme models to reduce the mismatch. Maximum likelihood eigenspace mapping [1] and maximum likelihood linear regression (MLLR) [2] have been considered for environment mismatch compensation.

Feature based approaches aim to find a transformation in the feature space to match an already trained model. Spectral subtraction (SS) along with many variants has been applied to this problem. For example, [3] applies non-linear SS to speech recognition in noisy car environment. In [4], cepstral mean normalization (CMN) and CDCN are applied to noisy car environments. All the above methods either require prior adaptation data from the test environment or process test data independent of the training environment to remove channel and noise related effects. Some methods exist in the literature which do not need separate adaptation data from the test environment and work on the test utterance directly. Feature transformation based on maximum likelihood framework calculates the additive bias vectors per test utterance which can be applied to incoming feature vector for mismatch compensation [5]. Subspace filtering based approaches ([6]) try to decompose the noisy signal in the time domain into orthogonal speech and noise subspaces assuming a low-rank linear model for speech and an uncorrelated additive noise. In case of correlated noise, noise pre-whitening transform is applied before the decomposition and a de-whitening transform is applied after it, but this also affects the original speech signal. This method focuses on finding the clean estimate of the speech signal and does not take into consideration the training environment statistics for mismatch reduction. Extensive research has also been focused on improving interactive systems for in-vehicle applications ([7], [8]). Noise modeling in the car environment based on *environmental sniffing* has been proposed by Akbacak and Hansen [9], and constrained switched adaptive beamforming for robust enhancement and recognition in car environment has also shown great improvements [10].

In this paper we propose a method based on the average eigenspace of speech and noise to reduce the mismatch between training and test environment. The remainder of the paper is or-

ganized as follows: Section 2 describes the concept of average eigenspace, Section 3 describes the application to current problem, Section 4 compares training and test environments through various statistics (other than WER obtained on the recognizer). Section 5 describes the baseline system and the results obtained. Section 6 discusses directions for future work and section 7 concludes the paper.

2. Average Eigenspace

We know that a set $\mathcal{R} = \{\mathbf{R}_k | k = 1, 2 \dots K\}$ of real symmetric matrices can be simultaneously diagonalized by a unitary transform if the matrices commute. Under this condition each matrix $\mathbf{R}_k$ in the set is similar to a diagonal matrix $\mathbf{\Lambda}_k$:

$$\mathbf{R}_k = \mathbf{U}\mathbf{\Lambda}_k\mathbf{U}^T, \qquad k = 1, 2 \dots K \tag{1}$$

where $\mathbf{U}^T$ is the unitary transform which diagonalizes all the matrices in the set. If the matrices do not commute, the transform needed to diagonalize the set is not unitary. However, we can attempt to obtain a unitary transform such that it makes the off-diagonal elements extremely small. One possible approach may be the minimization of the following criterion function with $\mathbf{U}$ being a unitary matrix:

$$f(\mathbf{U}) = \sum_{k=1,2\dots K} \sum_{\substack{1 \le i,j \le N \\ i \ne j}} \left| \left(\mathbf{U}^T \mathbf{R}_k \mathbf{U}\right)_{ij} \right|^2 \tag{2}$$

where $\left(\right)_{ij}$ denotes the (i, j) element of the matrix. The extended Jacobi technique for simultaneous diagonalization optimizes this criteria by iteratively applying plane rotations to all the matrices in the set $\mathcal{R}$ and minimizing the criteria for these rotations. The final transform is then calculated as the product of these plane rotations. A closed form expression for the optimal Jacobian angles for plane rotation is given by [11]. This transformation process is called approximate simultaneous diagonalization of the set $\mathcal{R}$, and $\mathbf{U}$ defines as we may call it - the *average eigenspace* of the matrix set $\mathcal{R}$. Every matrix $\mathbf{R}_k$ in the set is now similar to a matrix $\mathbf{\Lambda}'_k$ which is the *most diagonal* in a quadratic sense :

$$\mathbf{R}_k = \mathbf{U}\mathbf{\Lambda}'_k\mathbf{U}^T, \qquad k = 1, 2 \dots K \tag{3}$$

3. Average Eigenspace for Speech and Noise

If $\mathbf{R}_k$ in the above section are the covariance matrices of random vectors $\mathbf{r}_k$ for all $k = 1, 2, \dots K$, this transformation results in *almost* decorrelating the elements of each random vector. Let $\mathbf{K}_x$ and $\mathbf{K}_n$ denote the covariance matrix estimates of corrupted speech and noise respectively which are obtained as follows (let $\mathbf{x}$ and $\mathbf{n}$ be the corrupted speech and noise random vector respectively in the feature space, each of length N):

$$\begin{aligned} \mathbf{K}_x &= \frac{1}{N_x}\sum_{i=1}^{N_x} \mathbf{x}_i\mathbf{x}_i^T - \bar{\mathbf{x}}\bar{\mathbf{x}}^T \\ \mathbf{K}_n &= \frac{1}{N_n}\sum_{i=1}^{N_n} \mathbf{n}_i\mathbf{n}_i^T - \bar{\mathbf{n}}\bar{\mathbf{n}}^T \end{aligned} \tag{4}$$

where N_x and N_n are the total number of observations of corrupted speech and noise respectively, and $\bar{\mathbf{x}} = \frac{1}{N_x}\sum_{i=1}^{N_x} \mathbf{x}_i$ and $\bar{\mathbf{n}} = \frac{1}{N_n}\sum_{i=1}^{N_n} \mathbf{n}_i$. In our experiments, we use mel frequency cepstral coefficients (MFCC) as feature vectors. The joint diagonalization criterion for these two covariance matrices can be optimized using the extended Jacobian technique and we obtain an average eigenspace of speech and noise.

$$\begin{aligned} \mathbf{K}_x &= \mathbf{U}\mathbf{\Lambda}'_x\mathbf{U}^T \\ \mathbf{K}_n &= \mathbf{U}\mathbf{\Lambda}'_n\mathbf{U}^T \end{aligned} \tag{5}$$

We call the columns of the unitary transformation matrix $\mathbf{U}$ the *average eigenvectors*. Off-diagonal elements of $\mathbf{\Lambda}'_x$ and $\mathbf{\Lambda}'_n$ are very small (compared to the diagonal elements). The diagonal elements of these matrices give us the energies or variances for corrupted speech and noise in the feature space along these *average eigendirections*.

3.1. Average eigenspace for training environment

The covariance matrices of speech and noise ($\mathbf{K}_x^{tr}$ and $\mathbf{K}_n^{tr}$) are estimated from the training data using a voice activity detection (VAD) algorithm. Approximate joint diagonalization of these matrices gives the average eigenspace for the training environment ($\mathbf{U}^{tr}$). We compare the variance of speech and noise along each of the *average eigendirections* (we will call these *directions* from now on) and the information is retained only along those directions which have a high ratio of speech to noise variance. If $\mathbf{U}_p^{tr}$ is the matrix (of size $N \times P^{tr}$) constructed from the vectors corresponding to these directions (P^{tr} such directions), this is achieved by applying the following transform to the signal in the feature space (columns of $\mathbf{U}_p^{tr}$ span a subspace of the vector space spanned by columns of $\mathbf{U}^{tr}$):

$$\mathbf{x}_p = \mathbf{U}_p^{tr^T}\mathbf{x} \tag{6}$$

Note that this process is done for static, delta and delta-delta cepstrum separately. Covariance matrices of these feature streams are estimated and the average eigenspace is calculated for each separately. If the selected number of directions for static, delta and delta-delta cepstrum are $P_s^{tr}, P_d^{tr}, P_{dd}^{tr}$ respectively, then we obtain a complete feature vector of length $P_s^{tr} + P_d^{tr} + P_{dd}^{tr}$ (less than original feature vector length $3N$). The phoneme models are trained using these new feature vectors.

3.2. Average eigenspace for test environment

This approach is further applied during testing phase. Covariance matrices for speech and noise ($\mathbf{K}_x^{ts}$ and $\mathbf{K}_n^{ts}$) are estimated from the test data and their average eigenspace is computed ($\mathbf{U}^{ts}$). A transformation matrix is formed in which directions (columns of $\mathbf{U}^{ts}$) along which we have high ratio of speech to noise variance are retained (P^{ts} such directions) and the remaining columns are replaced with an all zero vectors. Let us call this matrix $\mathbf{U}_p^{ts}$ (having size $N \times N$ with $N - P^{ts}$ columns being zero). Original feature vectors are transformed using this matrix and we obtain a new feature vector $\mathbf{x}'_p$ of size $N \times 1$ with $N - P^{ts}$ entries as zero,

$$\mathbf{x}'_p = \mathbf{U}_p^{ts^T}\mathbf{x} \tag{7}$$

We need to transfer the information present in the feature vector $\mathbf{x}'_p$ to the restricted eigenspace of the training environment (restricted to selected directions, $\mathbf{U}_p^{tr}$). For this, we transfer the information in $\mathbf{x}'_p$ back to the MFCC space. The MFCC vector obtained is then transformed to the restricted space of the training environment. The final feature vector is given by,

$$\begin{aligned} \mathbf{x}_p &= \mathbf{U}_p^{tr^T}\mathbf{U}^{ts}\mathbf{x}'_p \\ &= \left(\mathbf{U}_p^{tr^T}\mathbf{U}^{ts}\mathbf{U}_p^{ts^T}\right)\mathbf{x} \end{aligned} \tag{8}$$

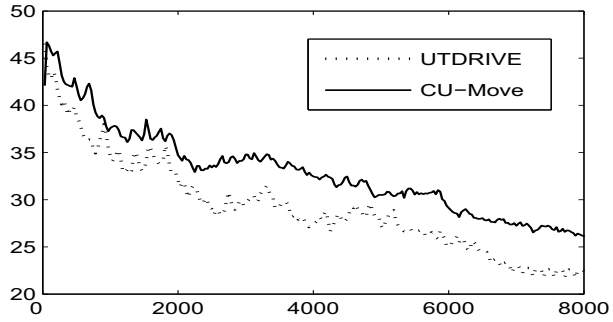

Figure 1: *Long term average noise spectra from UTDrive and CU-Move, obtained using 40 sec. duration of data (log magnitude (in dB) vs frequency (in Hz))*

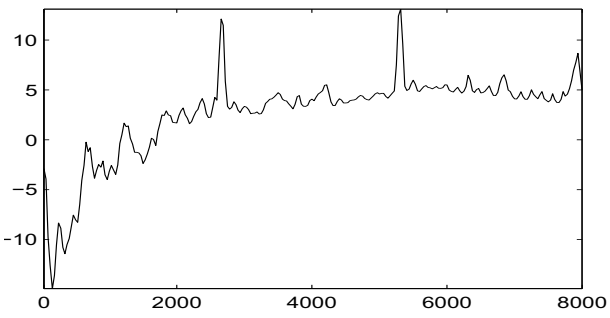

Figure 2: *Channel response applied to test data to incorporate convolutive mismatch in environments(log magnitude (in dB) vs frequency (in Hz))*

Again, this process is performed individually for static, delta and delta-delta cepstrum vectors. The final complete feature vector obtained has length $P_s^{tr} + P_d^{tr} + P_{dd}^{tr}$ (same as that in training).

4. Train and Test Environments

In our evaluations, we use two corpora collected from real car environments: UTDrive [12] and CU-Move [13]. The speech collected from a far field microphone is used in the experiments for both corpora. The acoustic conditions in both cars and microphones are different in the two databases. Here, we focus on quantifying the mismatch between these two environments.

To illustrate the distribution of noise across frequencies, we average the magnitude spectra of noise/silence frames over a duration of about 40 seconds. Fig. 1 shows the long term average noise spectra for both environments. Background noise in CU-Move is spread across a wider frequency range while noise in UTDrive drops significantly in the higher frequency range. It can be seen that average noise power in CU-Move is higher than UTDrive across the complete frequency band. There is a difference of 2-7 dB in the noise levels in the frequency range 0-4000 Hz.

Speech from close-talk microphone is also available as part of the UTDrive corpus. We try to estimate the channel mismatch between speech from the close-talk and far-field microphones. Long term average of the log spectra can be used to characterize the channel. Average log spectra of both channels are calculated and their difference is found (far-field log-spectra is subtracted from the close-talk log-spectra), which characterizes the channel mismatch, with the response shown in Fig. 2. This resembles a high pass filter and also has two sharp peaks in the spectrum. Far-field microphone speech from UTDrive is processed with this filter. This processing further accentuates the mismatch between training and test environments. Recognition experiments are performed with this data to test the effectiveness of the approach.

The Car environment presents additional challenges in the sense that background noise is non-stationary. There are many events that can happen which change the acoustic environment such as rolling up/down the windows, indicator beeps, etc. This varies the noise shape across the frequency range over time. The difference in environments is reflected in the word error rates, when training is done with one corpus and testing performed with the other corpus.

5. Recognition Experiments

In this section, we describe the results of large vocabulary continuous speech recognition (LV-CSR) experiments. The average eigenspace approach is used for preprocessing the speech in the feature space and recognizer is trained on these new features. Similarly during testing, the features are preprocessed using the average eigenspace transform for every frame before submitting to the recognizer.

5.1. Baseline

We use a speaker-independent LVCSR system based on CMU Sphinx for evaluating this approach. MFCCs in combination with their first and second order derivatives are used as the feature vector (39 dimensional feature vector). Acoustic modeling is done for a set of 42 phonemes. Each of the 127 phoneme states is modeled using a mixture of 8 Gaussian distributions without any tying (context-independent monophone models are used). A trigram language model for a 1k-word vocabulary is used during decoding for all experiments with CU-Move and UTDrive corpora.

5.2. Experiments and Results

We compute estimates of the covariance matrices for speech and noise as given by Eq. 4 for the training data using MFCCs as the feature vector. Average eigenspace is computed using extended Jacobian technique with plane rotation angles given by [11] with the minimizing criterion given by Eq. 2, and favorable *directions* (which have large speech information and low noise) are chosen. Training feature vectors are then transformed (Eq. 6) and the model trained using these features. In our experiments, CU-Move is used as the training corpus with a 23-dimensional feature vector after applying this transform (dimensionality is reduced from the original 39-D MFCC). During test, covariance statistics are collected again from the test data and the transform of Eq. 8 is applied to feature vectors before submission to the recognizer.

In all experiments, we collect the covariance statistics on a per utterance basis to estimate the average eigenspace. This means, a single transform is applied to the whole utterance (4-15 words per utterance). For the next utterance, this process of collecting covariance statistics, estimating the average eigenspace and finding the transform is repeated. It is also possible to estimate the average eigenspace more often (per word or per N_x frames), in which case the noise covariance statistics will be collected from the past N_n frames (to adapt to the dy-

Train with	Test on	Baseline (using 39-D MFCC)	CMN+SS (using 39-D MFCC)	Average Eigenspace (23-D features)
CU-Move	CU-Move	11.4%	6%	4.2%
	UTDrive	18.8%	19.5%	10.2%
	UTDrive_CH[1]	38.3%	36.1%	24.4%

[1] UTDrive speech processed with filter response shown in Fig. 2

Table 1: WER for baseline, SS+CMN and average eigenspace methods

namic nature of noise) and speech covariance statistics will be collected from the N_x frames for which the transform is to be estimated. Finding the transform more often is expected to give better results as it will give the average eigenspace, focusing on less number of speech frames, but at the same time requiring extensive computational resources.

Table 1 shows the results of the experiments. First, we train the ASR model using MFCCs without any preprocessing. For the matched environment case (train and test with CU-Move data), we get a WER of 11.4%. Applying cepstral mean normalization (CMN) and spectral subtraction reduces the WER to 6%. Application of the average eigenspace approach yields a WER of 4.2%. For the first mismatched case (train with CU-Move and test on UTDrive), the model is trained with unprocessed MFCCs producing a WER of 18.8%. Processing the speech with CMN and spectral subtraction (both during training and test) increases the WER in this case. The average eigenspace approach reduces the WER relatively by almost 45%. For the second mismatched case, where test is done on UTDrive data processed with the filter response shown in Fig. 2, the baseline WER is 38.3%. SS and CMN improve performance only marginally to 36.1%. The average eigenspace approach results in a reduction in WER to 24.4% (relative reduction of 36%).

6. Discussion

Having established the average eigenspace approach, there are a number of ways to consider applying the method to speech. In this paper, we focused on its application to MFCC feature vectors. Alternative features for recognition such as Mel frequency filter bank energies, PMVDR features, and others are also possible since the impact of noise will be different in each corresponding feature domain (additive, nonlinear, etc.).

In the experiments reported here, we estimate the average eigenspace on a per utterance basis during the test phase. The frequency of estimation of the average eigenspace could be explored, where we either increase the estimation rate, or update the estimation based on an additional metric of speech/acoustic/environment diversity over time. Increasing the rate of average eigenspace estimation is expected to improve performance, but at the cost of increased computational complexity. This trade-off can be adjusted for speech recognition depending on the application (e.g., near real-time for voice dialog applications, off-line ASR for transcript generation in spoken document retrieval).

7. Conclusions

In this paper, we introduced an approach for training and test environment mismatch compensation using the concept of average eigenspace. It is named as average eigenspace because a set of covariance matrices are *approximately* diagonalized along its principal axes. This approach is particularly effective for non-stationary environments such as in-vehicle applications. A side result of applying this approach is the reduction in dimensionality of the phoneme state output distributions. The approach was evaluated using two noisy car corpora - UTDrive and CU-Move. We obtain a significant reduction in WER for both mismatched as well as matched environmental conditions. The results suggest an effective means to address environment mismatch for robust speech recognition in time-varying vehicle conditions.

8. References

[1] P. Nguyen, C. Wellekens, and J. Junqua, "Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments," in *EUROSPEECH*, vol. 6, Sep. 1999, pp. 2519–2522.

[2] C. J. Legetter 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.

[3] P. Lockwood, J. Boudy, and M. Blanchet, "Non-linear spectral subtraction (NSS) and hidden markov models for robust speech recognition in car noise environments," in *IEEE ICASSP*, 1992, pp. 265–268.

[4] N. Hanai and R. M. Stern, "Robust speech recognition in the automobile," in *Proceedings of the International Conference on Spoken Language Processing*, Sep. 1994.

[5] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 3, pp. 190–202, May 1996.

[6] K. Hermus and P. Wambacq, "Assessment of signal subspace based speech enhancement for noise robust speech recognition," in *IEEE ICASSP*, vol. 1, May 2004, pp. 945–948.

[7] H. Abut, J. H. L. Hansen, and K. Takeda, *DSP for In-Vehicle and Mobile Systems*. Springer-Verlag Publishing, 2004.

[8] ——, *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards*. Springer Publishing, 2006.

[9] M. Akbacak and J. H. L. Hansen, "Environmental sniffing: Noise knowledge estimation for robust speech systems," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 465–477, Feb. 2007.

[10] X. Zhang and J. H. L. Hansen, "CSA-BF: A constrained switched adaptive beamformer for speech enhancement and recognition in real car environments," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 733–745, Nov. 2003.

[11] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM Journal of Matrix Analysis and Application*, vol. 17, no. 1, pp. 161–164, Jan. 1996.

[12] P. Angkititrakul and J. H. L. Hansen, "UTDrive: The smart vehicle project," in *In-Vehicle and Mobile Systems*. Springer Publishing, 2008, ch. 5.

[13] J. H. L. Hansen, X. X. Zhang, M. Akbacak, U. Yapanel, B. Pellom, W. Ward, and P. Angkititrakul, "CU-MOVE: Advanced in-vehicle speech systems for route navigation," in *DSP for In-Vehicle and Mobile Systems*. Springer-Verlag Publishing, 2004, ch. 2.