A KERNEL MEAN MATCHING APPROACH FOR ENVIRONMENT MISMATCH COMPENSATION IN 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

ABSTRACT

The mismatch between training and test environmental conditions presents a challenge to speech recognition systems. In this paper, we investigate an approach for matching the distributions of training and test data in the feature space. This approach uses the property of reproducing kernel Hilbert space (RKHS) with a universal kernel for the task of distribution matching. The approach is unsupervised, requiring no transcripts of data for compensation, and can be employed either with explicit adaptation data or with live test data. The approach is evaluated on two real car environments - CU-Move and UTDrive. Relative improvements of between 10-25% are obtained for different experimental setups.

Index Terms— speech recognition, feature adaptation, reproducing kernel hilbert space, universal kernel

1. INTRODUCTION

The performance of Automatic Speech Recognition (ASR) systems suffer dramatically when there is a mismatch between 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. For in-vehicle systems, there is a need to perform recognition using far-field or hands-free microphone so that the driver is free from unnecessary distractions.

Considerable research has been conducted on the problem of mismatch compensation for speech recognition. Some of these methods require the presence of stereo data (simultaneous recording from both environments) that 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. Most of them need prior adaptation data to do it.

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, [1] applies non-linear SS to speech recognition in noisy car environments. In [2], cepstral mean normalization (CMN) and (Codebook dependent cepstral normalization) CDCN are applied to noisy car environments. All these 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 has been proposed in [3]. [4] attempts to match the histograms of training and test data for normalization. Extensive research has also been focused on improving interactive systems for in-vehicle applications ([5], [6]).

Recently, Huang, et. al [7] applied a kernel mean matching approach to the problem of correcting sample selection bias (or domain adaptation) using unlabeled data. The solution of kernel mean matching (KMM) gives the weights for each training sample that are used to scale the loss function for each pair of (feature vector, class label). The weights obtained are the ratio of a test to training distribution under ideal asymptotic conditions (provided some constraints on the support of the distributions are met). The expectation of the weighted loss function is minimized to get the classifier parameters. The classifier should be optimal for the test distribution according to the theory of importance sampling.

However, it is well known that training a recognizer for speech is computationally complex, and once trained it is not feasible to retrain for the test distribution if the test distribution itself is not stationary (the noise and channel characteristics are changing). In this paper, we achieve compensation in the feature space using the kernel mean matching approach and assess performance for the task of speech recogntion. The remainder of the paper is organized as follows: Sec. 2 describes background concerning kernel mean matching; Sec. 3 describes the actual method employed; Sec. 4 describes the training and test environments and compares them using various statistics (other than WER obtained on the recognizer); Sec. 5 describes the baseline system and results obtained. Sec. 6 summarizes the findings of the paper.

2. BACKGROUND

Let $(\mathcal{X}, d)$ be a metric space and (p, q) be two probability measures defined on the Borel $\sigma-$algebra on $\mathcal{X}$. We have the following result from [8],

Lemma 1 *Two probability measures will be equal ($p = q$) if and only if $\mathbf{E_{x\sim p}(f(x))} = \mathbf{E_{x\sim q}(f(x))}$ for all $f \in C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of continuous bounded functions on $\mathcal{X}$.*

In practice, it is not feasible to work with such a rich class of functions. Alternatively, consider a restricted function class $\mathcal{F}$ and

This project was supported in part by USAF under a subcontract to RADC, Inc. under FA8750-05-C-0029. Approved for public release; distribution unlimited.

a metric for closeness/separability (integral probability metric [9], maximum mean discrepancy (MMD) [10]),

$$\mathrm{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} (\mathbf{E_{x \sim p}}[\mathbf{f}(\mathbf{x})] - \mathbf{E_{x \sim q}}[\mathbf{f}(\mathbf{x})]) \tag{1}$$

The computation of MMD becomes straightforward if we are in a reproducing kernel Hilbert space (RKHS). To define RKHS, let us take $\mathbf{X}$ to be an arbitrary set and $\mathbf{H}$ to be a Hilbert space of functions on $\mathbf{X}$. We say that $\mathbf{H}$ is a RKHS if the linear map $f \mapsto f(x)$ is continuous for all x in $\mathbf{X}$. This essentially means that for every x in $\mathbf{X}$, there exists an element K_x of $\mathbf{H}$ such that

$$f(x) = \langle f(\cdot), K_x(\cdot) \rangle \qquad \forall f \in \mathbf{H}.$$

Here, K_x is called the point evaluation functional at x. We can define a function $K : X \times X \mapsto \mathbb{R}$ by $K(x, y) := K_x(y)$. This is called the reproducing kernel for the Hilbert space H. For example, $\delta(x, \cdot)$ can be considered a point evaluation functional for L^2 (which is not a RKHS), but $\delta(x, \cdot)$ is not in L^2 (which is consistent with the fact that L^2 space is not RKHS). The reproducing property of the kernel says that $K(x, y) = \langle K(\cdot, x), K(\cdot, y) \rangle$. The map $\Phi : x \to K(\cdot, x)$ is called the feature map which maps each point in the domain of H to the function $K(\cdot, x)$ in the RKHS.

An RKHS with a universal kernel has a universal approximating property: given any positive number ϵ and any function $f \in C(X)$, there exists a function $g \in K(X)$ such that $\|f - g\|_\infty \leq \epsilon$ ([11]). That is, the universal RKHS is dense in $C(X)$.

It can be shown ([10]) that if $\mathcal{F}$ is a unit ball ($\mathcal{F} = \{f : \|f\| \leq 1\}$) in a universal RKHS defined on the compact metric space $\mathcal{X}$, then

$$\mathrm{MMD}[\mathcal{F}, p, q] = 0 \quad \text{if and only if} \quad p = q. \tag{2}$$

Let us define (where feature map $\Phi : x \to K(\cdot, x)$), $\mu(p) := \mathbf{E_{x \sim p}}[\mathbf{\Phi}(\mathbf{x})]$. It can be easily shown for a universal RKHS that ([10])

$$\mathrm{MMD}[\mathcal{F}, p, q] = \|\mu(p) - \mu(q)\|. \tag{3}$$

The consequence of these results (Eq. (2) and Eq. (3)) is that we can mimimize the separation between two distributions by minimizing Eq. (3). The fact that we are operating in RKHS makes the minimization computationally feasible.

3. KERNEL MEAN MATCHING FOR SPEECH

Let us suppose that training and test data is generated by two different distributions p and q respectively. Here, we match the two distributions by assuming an additive bias exists in the cepstral domain, although there can be other possible transforms to model the mismatch (like a linear transformation in the original feature domain). Formally,

$$\mathrm{MMD}^2[\mathcal{F}, p, q] = \|\mathbf{E_{x \sim p}}[\mathbf{\Phi}(\mathbf{x})] - \mathbf{E_{x \sim q}}[\mathbf{\Phi}(\mathbf{x})]\|^{\mathbf{2}}. \tag{4}$$

Let $\mathbf{x_i}(1 \leq i \leq n)$ and $\mathbf{x'_i}(1 \leq i \leq n')$ be the sequences of MFCC vectors for training and test speech respectively. Replacing the expectations by the empirical means and adding a bias term to the test MFCC vectors (before taking means in RKHS), the objective function $O(\beta)$ to be minimized becomes

$$O(\beta) := \Big\| f(\cdot) \Big\|^2 = \Big\| \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathbf{x_i}) - \frac{1}{n'} \sum_{i=1}^{n'} \Phi(\mathbf{x'_i} + \beta) \Big\|^2. \tag{5}$$

Simplifying the objective function,

$$\begin{aligned} O(\beta) =& \langle f(\cdot), f(\cdot) \rangle \\ =& \frac{1}{n^2} \Big\langle \sum_{i=1}^{n} \Phi(x_i), \sum_{j=1}^{n} \Phi(x_j) \Big\rangle \quad - \\ & \frac{2}{nn'} \Big\langle \sum_{i=1}^{n} \Phi(x_i), \sum_{j=1}^{n'} \Phi(x'_j + \beta) \Big\rangle \quad + \\ & \frac{1}{n'^2} \Big\langle \sum_{i=1}^{n'} \Phi(x'_i + \beta), \sum_{j=1}^{n'} \Phi(x'_j + \beta) \Big\rangle. \end{aligned} \tag{6}$$

Using the reproducing property of RKHS and omitting unnecessary terms,

$$\begin{aligned} \arg\min_\beta O(\beta) = \arg\min_\beta \Big(-\frac{2}{nn'} \sum_{i=1}^{n} \sum_{j=1}^{n'} K(x_i, x'_j + \beta) \quad + \\ \frac{1}{n'^2} \sum_{i=1}^{n'} \sum_{j=1}^{n'} K(x'_i + \beta, x'_j + \beta) \Big). \end{aligned} \tag{7}$$

The Gaussian kernel is universal on every compact subset of $\mathbb{R}^n$ ([11]), and we employ a Gaussian kernel in our experiments. As different phoneme classes in the speech signal have significantly different distributions, it is reasonable to restrict the optimization problem to a subset of train and test samples which have high probability of belonging to the same class. The objective function is then given by

$$\begin{aligned} &\arg\min_\beta O(\beta) = \\ &\arg\min_\beta \Big(-\frac{2}{nn'} \sum_{i=1}^{n'} \sum_{j=1}^{n'} \exp(-\sigma^2 \|x_i - x'_j - \beta\|^2) \\ &\quad + \frac{1}{n'^2} \sum_{i=1}^{n'} \sum_{j=1}^{n'} \exp(-\sigma^2 \|x'_i - x'_j\|^2) \Big) \\ &\qquad \text{(if } \beta \text{ is same for all } j) \\ &= \arg\min_\beta \Big(-\frac{2}{nn'} \sum_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n' \\ \|x_i - x'_j\| < D}} \exp(-\sigma^2 \|x_i - x'_j - \beta\|^2) \Big) \end{aligned} \tag{8}$$

where $\mathbf{x_i}$ is a sample from training data and $\mathbf{x'_j}$ is a sample from test data. The value of σ influences the rate of decrease in the dot product in kernel space with the increase in Euclidian distance in the original feature domain. A greater value of σ means that the dot product decreases more rapidly with increasing Euclidian distance between the two feature vectors in the original space. The second term in the equation (7) cannot be removed if we choose to use a different β for each frame. Although in this paper we experiment with a fix beta for a set of test samples (one utterance at a time), the possibility of a different β for each frame or group of frames can also be explored by employing a suitable constraint on the values of β. We also experiment with taking different β for speech and silence parts.

The optimization is limited to inside a ball of radius D centered at each test sample. For a fixed β for all frames, the method reduces to a simple form as in Eq. 8. This equation in itself can be explained intuitively; however it is interesting to have it coming from

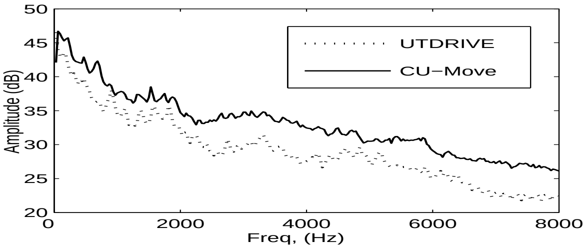

Fig. 1. *Long term average noise spectra from UTDrive and CU-Move, obtained using 40 sec. duration of data (log magnitude (in dB) vs frequency (in Hz))*

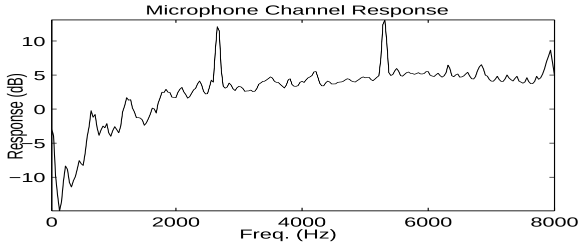

Fig. 2. *Channel response applied to test data to incorporate convolutive mismatch in environments(log magnitude (in dB) vs frequency (in Hz))*

a distribution matching viewpoint using properties of RKHS with a universal kernel. Here, a gradient descent is used to minimize the function. Let us concentrate on Eq. (8) for a moment and note the differences of this with standard CMN, which also employs an additive bias. The bias in this method is obtained by taking into account both training and test data and minimizing the mismatch. CMN considers one speech segment at a time and obtains the bias based on that. It simply tries to align the means of the training and test speech utterances to zero. The proposed method tries to match the distributions by matching training and test data means in the kernel space. The complexity of this method is higher than CMN because of pairwise computations but it is reduced to a great extent by restricting the optimization among the samples of same classes. Finally, although we have matched distributions by an additive bias here, an affine transform operating on test data can also be considered using the same framework.

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 only the 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 the average noise power in CU-Move is higher than UTDrive across the complete frequency band. There is a difference of 2-7 dB in noise levels over the frequency range 0-4000 Hz.

Speech from a close-talk microphone is also available as part of the UTDrive corpus. An estimate is also made of the channel mismatch between speech from the close-talk and far-field microphones. The long term average of the log spectra can be used to characterize the channel. The 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 from far-field to close-talk, with the response shown in Fig. 2. This resembles a high pass filter with two sharp peaks in the spectrum. These peaks are due to some measurement artifact in some of the close-talk speech files. In these experiments, far-field microphone speech from UTDrive is processed with this filter to make it closer to the close-talk channel, which further increases the mismatch between training and test environments. Recognition experiments are performed with this data to test the effectiveness of the feature based environment mismatch compensation approach.

The two car environments present additional challenges in the sense that each 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, road surface, engine noise, etc. This varies the noise shape across the frequency range over time. The difference in environments is reflected in word error rates, when training is done with one corpus and testing performed with the second corpus.

5. RECOGNITION EXPERIMENTS

5.1. Baseline

We use a speaker-independent ASR system based on CMU Sphinx for evaluating the proposed approach. Mel frequency cepstral coefficients (MFCC) 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 state tying (context-independent monophone models are used). A 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

Typical utterances, that have high acoustic scores on the trained ASR model, are chosen from the training data. Samples from these typical utterances are used as training distribution samples in the optimization problem. The values of σ and D in these experiments is taken to be 0.2 and 1 (both values are decided empirically). Two types of experiments are conducted. First, the bias vectors on a per utterance basis are estimated. Next, we estimate different bias vectors for the speech and silence parts. This is done to acknowledge the fact that speech has a significant impact of channel in addition to additive noise while silence segments are mainly corrupted by additive noise. For this, the training samples are grouped in two classes - speech and silence. Voice activity detection is performed on the test utterance to decide its class and two different optimizations are carried out.

Training is done on the data from CU-Move (far-field), while testing is done with data from CU-Move far-field microphone (matched condition), UTDrive far-field data and UTDrive far-field data processed with a close-talk channel as shown in Fig. 2 (mismatched conditions). Table 1 shows the results obtained. Cepstral mean normalization (CMN) tries to align the means of all training and test utterances to zero. This is expected to reduce the channel mismatch to some extent. As the data used in the experiments are from noisy car environment, we also use spectral subtraction (SS)

Train with CU-Move	Test On		
	CU-Move	UTDrive	UTDrive_CH[1]
Baseline	13.8%	23.1%	38.4%
CMN+SS	9.1%	22.4%	34.3%
CMN+SS+RATZ	9.4%	22.8%	34.0%
CMN+SS+KMM	8.8%	21.7%	29.5%
CMN+SS+KMM (silsp[2])	8.6%	21.3%	29.2%

[1] UTDrive speech processed with filter response shown in Fig. 2
[2] different bias vectors for silence and speech segments

Table 1. Word Error Rate (WER) results

which tries to suppress the additive background noise and works in the spectral domain. It can be seen that kernel mean matching improves the performance in all the three experiment setups. Using different biases for silence and speech improves performance marginally. This can be due to the fact that we have already reduced the possibility of optimization across classes by employing the constraint as in Eq. (8). The most improvement (about 25% relative) is seen in the heavily mismatched third case. Feature compensation using RATZ [14] is also applied in all the three test scenarios. RATZ tries to compute the correction factors which can be applied to the parameters of training distribution to get an estimate of the parameters of test data distributions. It tries to find the MMSE estimate of training speech given the test speech. As the closed form solution is hard to get, it models the effect of environment as an additive term in the cepstral domain. It uses the correction factors for means as these additive terms. CUMove data is used to train a gaussian mixture model of 256 components and complete test data is used for correction factor estimation. Application of RATZ gives marginal improvement in the third case (compared to CMN+SS) while the performance actually degrades in the other two cases.

For the matched environment case, the baseline WER is 13.8%. Applying spectral substraction (SS) and cepstral mean normalization (CMN) reduces it to 9.1%. Application of kernel mean matching yields a WER of 8.8%. Use of different bias vectors for silence and speech reduces it further to 8.6%. For the first mismatched condition (test on UTDrive data), the baseline WER is 23.1%, CMN and SS reduce it to 22.4%. The proposed approach reduces the WER to 21.7% (single bias for the whole utterance) and 21.3% (different bias for speech and silence). For the second mismatched case (test on UTDrive data processed with close-talk channel response), the baseline WER is 38.4%. Application of CMN and SS gives a WER of 34.3%. The proposed approach reduces the WER to 29.5% (single bias) and 29.2% (different biases for speech and silence segments), which is a relative improvement of 24% with respect to the baseline.

6. CONCLUSION

A new approach to compute bias vectors using kernel mean matching was presented for feature based mismatch compensation. Although the mismatch in this paper is modeled using an additive bias in the cepstral domain, we can also use a linear transformation in addition to the bias term to model the mismatch. The same approach can be used without any modification to estimate both these factors. The bias vectors are estimated per utterance. This can be extended for estimating biases per frame or a group of homogeneous frames with some suitable constraints on the optimization problem so that the biases do not transfer the frames into a different phoneme class. One bias per utterance ties all the frames together which are moved in a single direction, so the prospect of a final feature frame moving to an undesignated region is negligible. There is a considerable improvement in the performance using the proposed approach. Still, there might be several factors affecting it. First, the objective function is highly non-linear and the solution can be trapped at a local minimum depending upon the initialization. Second, different phonemes in speech have significantly different distributions and the use of the Euclidean distance to ensure that the optimization is over the same class is not as sophisticated an approach. Some probabilistic method can be used here. Third, an additive bias alone may not be able to model the entire mismatch between environments. Some kind of non-linear term (which is a function of the input feature itself) or an affine transformation in addition to the bias may give better results. The prime purpose of this paper is to consider the problem of environment mismatch from the viewpoint of a difference in probability distributions of the feature vectors. To that end, the formulation and real car data evaluations show promise for in-vehicle ASR.

7. REFERENCES

[1] 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.

[2] N. Hanai and R. M. Stern, “Robust speech recognition in the automobile,” in *ICSLP*, 1994.

[3] A. Sankar and C.-H. Lee, “A maximum-likelihood approach to stochastic matching for robust speech recognition,” *IEEE Trans. Speech Audio Proc.*, **4***(3)*, pp. 190–202, 1996.

[4] S. Molau, D. Keysers, and H. Ney, “Matching training and test data distributions for robust speech recognition,” *Speech Comm.*, 2003.

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

[6] M. Akbacak and J. H. L. Hansen, “Environmental sniffing: Noise knowledge estimation for robust speech systems,” *IEEE Trans. Audio, Speech and Language Proc.*, **15***(2)*, 2007.

[7] J. Huang, A.J. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf, “Correcting sample selection bias by unlabeled data,” in *NIPS*. 2007, MIT Press.

[8] R. M. Dudley, *Real Analysis and Probability*, Cambridge University Press, Cambridge, UK, 2002.

[9] A. Muller, “Integral probability metrics and their generating class of functions,” *Adv. Appl. Prob.*, vol. 29, 1997.

[10] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schlkopf, and A. J. Smola, “A kernel method for the two-sample-problem,” in *NIPS*. 2006, MIT Press.

[11] I. Steinwart, “On the influence of the kernel on the consistency of support vector machines,” *Journal of Machine Learning Research*, vol. 2, pp. 67–93, 2001.

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

[13] J. H. L. Hansen et al., “CU-MOVE: Advanced in-vehicle speech systems for route navigation,” in *DSP for In-Vehicle and Mobile Systems*, chapter 2. Springer Publishing, 2004.

[14] P.J.Moreno, B.Raj, and R.M. Stern, “Data-driven environmental compensation for speech recognition: A unified approach,” *Speech Comm.*, vol. 24, pp. 267–285, 1998.