A PARALLEL ENVIRONMENT MODEL (PEM) FOR SPEECH RECOGNITION AND ADAPTATION

Mazin Rahim

AT&T Labs - Research, 180 Park Avenue, Florham Park, NJ 07932, USA
mazin@research.att.com

ABSTRACT

A speech recognition system for modeling an acoustic mismatch across different environments is presented. The basic philosophy is to apply discriminative learning techniques to separate the *recognition process*, that is represented by a hidden Markov model (HMM), from the *environmental process* which is denoted by a limited number of translation vectors. Each segment of speech is assigned to an environment and recognition is performed upon projecting the parameters of the HMM to best characterize the acoustic space of that environment. The proposed system provides an interesting framework for better modeling and adaptation of speech signals with varying acoustic conditions. Experimental findings on connected digits recognition for three different environments are reported.

1. Introduction

One classical problem with automatic speech recognition (ASR) systems is their ability to maintain *robustness* to a large variety of mismatched acoustic conditions that exist between the training model and the testing environment. A family of techniques have been proposed during the past several years for dealing with this type of problem, ranging from feature to model compensation methods. In essence, the objective is to transform the features of an ASR system in such a way that would bring them within the vicinity of the statistical model space and/or to transform the model parameters in order to better characterize the distorted feature space [3]. Lines of research in this area include the family of cepstral normalization techniques by CMU [9], bias removal [7], stochastic matching [8], parallel model combination [1], etc. By assuming some naive model of the mismatch, these techniques can successfully compensate for signal distortion resulting, generally, in a moderate improvement in recognition performance. Clearly, the inaccurate assumptions in modeling the mismatch limit the capabilities of these compensation techniques especially when dealing with real-world situations such as different network conditions, transducers, speakers, noise, etc.

To enhance robustness particularly when deploying a speech recognition service, it is customary to train the statistical model (e.g., HMM) on a wide range of speech data in the hope of learning and maintaining robustness across a spectrum of acoustic conditions. Although training on a variety of data collections is known to improve ASR performance, the resultant "diffused" model with its large variances become unsuitable for achieving high-accuracy in any one particular environment. As pointed out by Lee [3], being able to cope with large training data in an efficient manner could potentially provide more precise acoustic models for speech recognition. Current state-of-the-art ASR systems employ an integrated strategy which uses large training data as well as feature/model compensation techniques in order to achieve competitive recognition performance.

In this paper, we propose a parallel environment model (PEM) for characterizing an acoustic mismatch across different environments. PEM is similar in spirit to the environment-independent cepstral compensation technique [9] and to the framework of family discovery [5] in that training is performed on each environment separately. However, the basic philosophy behind PEM is to use discriminative learning techniques to isolate the *recognition process*, that is represented by an HMM, from the *environmental process* which is denoted by a set of *translation vectors*. The HMM conducts the basic modeling of the speech units while the translation vectors help to transform the HMM into an acoustic space that is more appropriate for the testing environment. In principle, this framework should lead to an improved recognition system for better modeling and adaptation of speech signals with varying acoustic conditions.

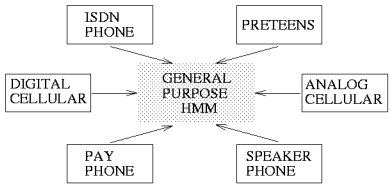

Figure 1: A schematic diagram of the PEM system

Figure 1 shows a schematic diagram of the PEM system which includes a general purpose HMM, referred to as the *base* model, along with a family of parallel models, each representing a different environment. An environment may refer to a collection of data that is spoken by specific types of users (e.g., preteens), or recorded over particular network conditions (e.g., digital cellular). An environment may also be considered as a subset of the training data that can be partitioned by maximizing some optimality criterion (e.g., likelihood) as will be illustrated in Section 2.1. Speech recognition using the PEM system involves identifying the most appropriate environment for each testing utterance and then transforming the parameters of the base model to best characterize the acoustic space of that environment. Model transformation involves shifting the Gaussian parameters using a limited number of translation vectors that are discriminatively trained on each environment separately. Due to the small number of these vectors, PEM is capable of handling several environments at a relatively small computational cost. Further, the parallel architecture of the PEM system provides a suitable framework for long-term adaptation of the model

parameters as will be illustrated in Section 2.3.

2. THE PEM SYSTEM

The PEM system consists of a HMM (a base model), Λ_0, representing the recognition process and a set of M parallel models, $\{\Lambda_l\}_{l=1,M}$, representing the environmental process. The objective of this system is to provide an improved framework for coping with acoustic variations that are caused by multiple recording conditions. The basic strategy is to transform the parameters of the base model to best characterize the acoustic space of the testing data. Parameter transformation is performed using a limited number of discriminatively-trained translation vectors. In this section, we describe the method for creating and identifying an environment, training and adapting the translation vectors and, finally, evaluating the PEM system.

2.1. Environment Classification

Let $X = \{x_1, x_2, .., x_T\}$ be a test sequence of T feature vectors, and $\{C_l\}_{l=1,M}$ be a set of M classes, or environments. When performing environment classification, the task is to identify the environment that *best* characterizes the feature space of X. If $p(X, C_l)$ is the joint probability density function for X and C_l, then the optimal classifier is the one that satisfies the decision rule

$$C_{l^*} = \arg\max_l P(X, C_l). \tag{1}$$

A classifier may be represented by a vector quantization (VQ) codebook, a Gaussian mixture model (GMM), a HMM or a neural network. A study using a VQ codebook and a GMM, both adopting the same number of parameters, showed that the two classifiers are equivalent in terms of accuracy. Accordingly, to minimize computational effort, PEM employs a set of VQ codebooks $\{\mathcal{C}_l\}_{l=1,M}$, such that the environment with a codebook that minimizes the distortion measure

$$\mathcal{C}_{l^*} = \arg\min_l d(X, \mathcal{C}_l) \tag{2}$$

is selected. $d(X, \mathcal{C}_l)$ is essentially a weighted Euclidean distance.

The knowledge of associating X with a particular environment during the training phase can be either available or established based on some optimality criterion. For example, if multiple data collections are utilized in training, each collection may then be considered as a separate environment. Consequently, a VQ classifier can be designed for each environment by applying, for example, the Lloyd clustering algorithm. Alternatively, if knowledge of the environment is unavailable, we may adopt matrix quantization in which clustering is applied at the utterance level, rather than at the frame level. The objective would be to build a set of codebooks $\{\mathcal{C}_l\}$ that minimizes the overall distortion, D, over, say, N training utterances:

$$D = \sum_{i=1}^{N} d(X_i, \mathcal{C}_{l^*}) \cdot 1(X_i \in \mathcal{C}_{l^*}), \tag{3}$$

where $1(\cdot)$ is an indicator function and $d(X_i, \mathcal{C}_{l^*})$ is a weighted Euclidean distance between the features X_i and the classifier corresponding to the nearest environment l^*.

2.2. Model Training

In principle, separating the recognition process from the environmental process provides an interesting framework that could facilitate a better understanding of the acoustic mismatch in speech recognition. In PEM, this problem is defined as finding an environment-specific transformation, $\mathcal{M}_{\Psi_l}$, with corresponding set of parameters Ψ, that satisfies $\Lambda_l = \mathcal{M}_{\Psi_l}(\Lambda_0)$. In this study, $\mathcal{M}_{\Psi_l}$ is represented by a limited number of translation vectors that are discriminatively-trained on each environment separately.

We consider model transformation for an environment l as a deviation in the parameters of the base model, Λ_0, from those in the target environment, Λ_l. Thus $\Lambda_l = \Lambda_0 + \Delta\Lambda_l$, where $\Delta\Lambda_l$ is referred to as the environment process for l. In PEM, $\{\Delta\Lambda_l\}$ are computed using the minimum classification error (MCE) framework that was proposed by Juang and Katagiri [2]. Given a speech token (or utterance), X_i, for class string i that belongs to environment l, discriminative training of the parameters of Λ_l is performed by minimizing the following loss function over the entire training data:

$$J = \sum_i \mathcal{S}\{d(X_i, \Lambda_0)\} \cdot 1(X_i \in \mathcal{C}_l), \tag{4}$$

where $\mathcal{S}\{\cdot\}$ is a sigmoid non-linear activation function and $d(\cdot)$ is a *misclassification* measure which is essentially a normalized log likelihood ratio between the correct string hypothesis and alternative (competing) hypotheses to string class i (see details in [2]).

The objective of MCE training, as applied to PEM, is to estimate the parameters of the environmental process, $\Delta\Lambda_l$, that minimize the loss function J. This is achieved through gradient descent, such that at the n^{th} iteration,

$$\Delta\Lambda_l^{(n)} = -\epsilon_n \frac{\partial J}{\partial \Lambda}\Big|_{\Lambda_l = \Lambda_l^{(n-1)}}, \quad \epsilon_n > 0, \ \Lambda_l^{(0)} = \Lambda_0. \tag{5}$$

When the acoustic mismatch between the base model and a given environment is not severe, it is expected that some of the parameters of Λ_0 will undergo negligible change. Further, training Λ_0 on each environment separately may lead some parameters to evolve in the same manner, thus demonstrating similar learning characteristics. To design a more efficient PEM system, we introduce *tying* among model parameters that display similar learning (or no learning) behavior. In essence, we cluster the parameters of $\Delta\Lambda_l$ for each environment into a limited number of translation vectors, $\{\mathcal{V}_l\}$. The resultant model $\hat{\Lambda}_l = \Lambda_0 + \mathcal{V}_l$ is further trained by minimizing the loss function in Eqn. 4 with respect to the parameters of $\hat{\Lambda}_l$, as opposed to Λ_0. This notion of clustering the translation vectors based on their learning characteristics rather than their acoustic information is similar to that proposed in vector field smoothing [4], and can be considered as a form of parameter tying. The nature of this tying is particularly appropriate for discriminative training since, unlike conventional tying schemes, it avoids integrating similar acoustic parameters that may correspond to confuseable events. Tying can potentially reduce the size of the environment models by a factor of ten causing only a minor effect on the recognition performance (see Section 4).

2.3. Long-term Adaptation

In telephone speech recognition, conventional systems that rely on a single HMM are known to perform poorly when attempting to do long-term adaptation on a wide-range of acoustic conditions. The parallel architecture of the PEM system, however, provides a more suitable

framework for adaptation and can potentially lead to a significant improvement in recognition performance. Further, the compact representation of the environment models through tying can provide an additional benefit when performing adaptation with a limited set of data.

In PEM, long-term adaptation affects solely the parameters of the desired environment model (i.e., the translation vectors). Given a set of adaptation data, $\{Y_j\}$, we optimize the translation vectors $\{\mathcal{V}_l\}$ in order to minimize the loss function

$$J_a = \sum_j \mathcal{S}\{d(Y_j, \hat{\Lambda}_l)\} \cdot 1(Y_j \in \mathcal{C}_l). \quad (6)$$

This procedure is a straight-forward extension to the training process described in Section 2.2.

2.4. Evaluation of the PEM system

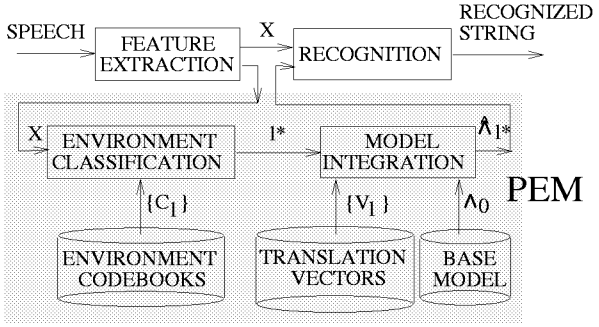

Figure 2: A block diagram of the PEM speech recognition system.

Fig. 2 shows a block diagram of the PEM speech recognition system. Following feature extraction of X, the most likely environment, l^*, which minimizes the distortion measure in Eqn. 2 is identified. Model integration is then performed which includes transforming the parameters of the base model Λ_0 to best characterize the acoustic space of X. The newly constructed HMM, namely $\hat{\Lambda}_{l^*} = \Lambda_0 + \mathcal{V}_{l^*}$, is then passed to the recognizer which performs standard Viterbi beam search.

This entire process of identifying an environment and transforming the base model requires less than a 3% of CPU time for each additional environment added to the PEM system.

3. DATABASE AND BASELINE PERFORMANCE

A speaker-independent telephone based connected digits database was used in this study. Utterances ranging from one to sixteen digits in length, were taken from three different field trial data collections. The first collection, referred to as "preteens", included children between the ages of 6 and 18 calling over a wireline network. The second collection, referred to as "adults-wireline", included adult speakers between the ages of 16 and 70 calling over a wireline network. The third collection, referred to as "adults-wireless" included adult speakers calling over a cellular network which was mostly an analog. In all three collections, data was recorded over a variety of environmental conditions and using various transducer equipments. The training and testing data included 5767 strings and 1261 strings for "preteens", respectively, 9562 strings and 818 strings for "adults-wireline", respectively, and 15487 strings and 1118 strings for "adults-wireless", respectively. Although each collection represented a wide range of acoustic variations it was considered as an *individual* environment in this study.

During feature extraction, a set of 39 features per frame was computed. This included 12 LPC-based cepstral coefficients plus a log energy along with their first and second order time derivatives. Each feature vector in the baseline system was directly passed to the recognizer which modeled each word (i.e., digit) in the vocabulary by a set of left-to-right continuous-density HMMs [6]. Each word was divided into three units, namely, head, body and tail. To model inter-word coarticulation, each word was made to have a single body with multiple heads and tails, resulting in a total of 274 sub-word models. Each sub-word model consisted of 3 to 4 states, with each state having a mixture of 4 Gaussian components.

System	ENV1	ENV2	ENV3	AVG
Global-ML	16.1	5.1	4.4	6.9
Global-MCE	8.1	2.4	1.3	3.0
Matched	6.2	1.9	0.8	2.3
PEM	6.6	1.9	0.9	2.4
PEM*	6.8	1.7	1.1	2.4
Adapt-Global	6.5	2.7	1.6	2.9
Adapt-PEM	5.5	2.0	1.0	2.3
Adapt-PEM*	6.3	1.8	1.2	2.4

Table 1: Percentage word error rate for various recognition systems. ENV1 denotes "preteens", ENV2 denotes "adults-wireline", ENV3 denotes "adults-wireless" and AVG denotes the average performance over the three environments.

Table 1 presents the performance of the baseline recognizer when applying maximum likelihood (ML) training on the entire three data collections (labeled as "Global-ML"). The results include the word error rate (including insertions, deletions and substitutions) when testing on each environment separately. To improve the baseline system performance, the Gaussian means of the HMMs were further optimized using MCE training. The results of this experiment are shown in Table 1 (labeled as "Global-MCE") which reflect our state-of-the-art performance when training on the entire data set. This amounts to a 56% reduction in the word error rate over the "Global-ML" system.

4. EXPERIMENTS USING PEM

The extent in which global training affects the recognition performance was further investigated. An experiment was conducted in which discriminative training, using the "Global-ML" model, and testing was done on each environment separately. The results of this experiment are shown in Table 1 (labeled as "Matched") which suggest that matched environment training and testing leads to a further drop in the word error rate by about 23% over the "Global-MCE" performance. This improvement supports the notion that global training on multiple environments could result in "diffused" recognition models.

In the next set of experiments, we evaluated the PEM recognition system in Fig. 2. Recall that the intent when incorporating PEM was to transform the base model parameters to better characterize the acoustic space of the

testing environment. The "Global-ML" was considered as the base model in this experiment and each environment was associated with a VQ classifier for identification and a set of translation vectors for model transformation. Each VQ classifier included 64 codewords and was designed based on cepstral information only. Translation vectors were assigned for all Gaussian means and were trained using the procedure outline in Section 2.2.

Table 1 presents the word error rates (labeled as "PEM") when evaluating the PEM system on the three selected environments. These results compare favorably with those reported when training and testing on the same environment (see results for "Matched"). Although environment classification in this experiment was only 91% correct, it is certain from the results that this did not play a major role in degrading the overall recognition performance.

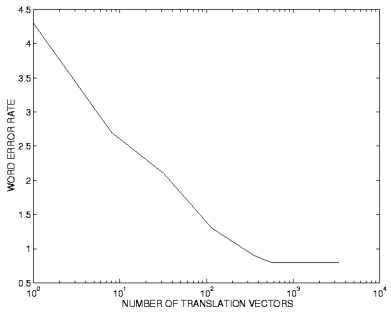

Figure 3: Word error rate as a function of the number of translation vectors.

The procedure for reducing the number of translation vectors, as described in Section 2.2, was evaluated in an experiment on the "adults-wireless" data. Fig. 3 shows the word error rate as a function of the number of translation vectors. These results, obtained by clustering the translation vectors and further training the set through MCE, demonstrate that the recognition performance can be maintained even when reducing the number of translation vectors by a factor of 10 (the initial model included 3352 Gaussian distributions). With this level of compression, we evaluated the PEM system with all the three environments active simultaneously. The results are given in Table 1 (labeled as "PEM*") which closely match those reported for "PEM".

Finally, we performed supervised adaptation, as described in Section 2.3, on three of the systems used, namely, "Global-MCE","PEM" and "PEM*". A subset of 94 strings were used, mostly taken from the "preteens" collection. From the results given in Table 1 we notice that all the three systems improved on the "preteens" environment. However, the conventional "Global-MCE" recognizer showed a moderate degradation in performance when tested on the other two environments. Since the models for "adults-wireline" and "adults-wireless" were not highly affected during adaptation, the performance on their respective data collections when using the "PEM" and "PEM*" systems was almost unchanged.

5. Summary

This paper presented a speech recognition system for modeling an acoustic mismatch across different recording environments. In the so called, parallel environment model (PEM), discriminative learning techniques were applied to isolate the environmental process from the basic modeling of speech units. The PEM system was shown to provide an efficient framework for better modeling and adaptation of speech signals with varying acoustic conditions. Our experimental results on connected digits recognition, corresponding to three different environments, demonstrated (a) a reduction in the word error rate by about 20% when using the PEM system over a traditional recognizer employing global training, (b) a minimal change in recognition performance when reducing the size of the environment models by a factor of 10, and (c) supervised adaptation resulted in a little change in performance for environments unseen during training. Experimenting with the PEM system using different number of mixture components, mixture tying and bias removal techniques is currently in progress.

6. REFERENCES

1. M. J. F. Gales and S. Young. An improved approach to hidden Markov model decomposition of speech and noise. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 223–226, 1992.
2. B.-H. Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 40:3043–3054, 1992.
3. C.-H. Lee. On feature and model compensation approach to robust speech recognition. In *Robust Speech Recognition for Unknown Communication Channels*, pages 45–54, 1997.
4. K. Ohkura, M. Sagiyama, and S. Sagayama. Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs. In *Proc. ICSLP '92*, pages 369–372, 1992.
5. S. M. Omohundro. Family discovery. In *NIPS '96*, pages 402–408, 1996.
6. R. Pieraccini and A. E. Rosenberg. Coarticulation models for continuous digit recognition. In *Proc. Acoust. Soc. Am.*, page 106, May 1990.
7. M. G. Rahim and B.-H. Juang. Signal bias removal by maximum likelihood estimation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(1):19–30, January 1996.
8. A. Sankar and C.-H. Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(3):190–202, May 1996.
9. R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima. *Signal Processing for robust speech recognition.* Chapter 15 in Automatic Speech and Speaker Recognition: Advanced Topics, C.H. Lee, F.K. Soong and K.K. Paliwal, Kluwer, Reading, MA, 1996.