ROBUST RECOGNITION OF CELLULAR TELEPHONE SPEECH BY ADAPTIVE VECTOR QUANTIZATION

M. Kemal Sönmez[1,2] *Raja Rajasekaran*[1] *John S. Baras*[2]

[1]Speech Research Laboratory, Systems and Information Sciences Laboratory, Texas Instruments, Inc., Dallas, TX
[2]Institute for Systems Research and Department of Electrical Engineering, University of Maryland College Park, MD

ABSTRACT

The performance degradation as a result of acoustical environment mismatch remains an important practical problem in speech recognition. The problem carries a greater significance in applications over telecommunication channels, especially with the wider use of personal communications systems such as cellular phones which invariably present challenging acoustical conditions. In this work, we introduce a vector quantization (VQ) based compensation technique which *both* makes use of a priori information about likely acoustical environments *and* adapts to the test environment to improve recognition. The technique is progressive and requires neither simultaneously recorded speech from the training and the testing environments nor EM-type batch iterations. Instead of using simultaneously recorded data, the integrity of the updated VQ codebooks with respect to acoustical classes is maintained by endowing the codebooks with a topology and using transformations which preserve the topology of the reference environment. We report results on the McCaw Cellular Corpus where the technique decreases the word error for continuous ten digit recognition of cellular hands free microphone speech with land line trained models from 23.8% to 13.6% and the speaker dependent voice calling sentence error from 16.5% to 10.6%.

1. INTRODUCTION

State of the art speech recognizers exhibit a particular sensitivity to mismatches in training and testing environments. This sensitivity degrades performance in many tasks such as command and digit recognition over telephone, voice dialing, etc. and is currently one of the most important practical problems in speech recognition. It carries a greater significance in applications over telecommunication channels, especially with the wider use of personal communications systems such as cellular phones which invariably present challenging acoustical conditions.

In this work, we describe an environment adaptation technique based on Adaptive VQ (AVQ) by topology preserving transformations. It utilizes a priori information about likely acoustical environments in the form of environment codebooks derived off-line from the reference environment codebook, *and* adapts on-line to the test environment to improve recognition. The technique requires neither simultaneously recorded speech from the training and the testing environments nor EM-type batch iterations. Instead of using stereo recorded data, the integrity of the updated VQ codebooks with respect to acoustical classes is maintained by endowing the codebooks with a topology and using transformations which preserve the topology of the reference environment.

The organization of the paper is as follows: Section 2 presents a brief review of VQ-class dependent compensation/adaptation techniques in the literature. The modeling of distortion as a difference error field and discrete approximations by VQ are included in Section 3. The approach we propose in the paper is developed in Section 4, and its results on the McCaw Cellular Corpus where it decreases the word error for continuous ten digit recognition of cellular hands free microphone speech with land line trained models from 23.8% to 13.6% and the speaker dependent voice calling sentence error from 16.5% to 10.6% are in Section 5.

2. RELATED PRIOR WORK

Speech recognition in noisy environments is an important practical problem and has attracted significant amount of research. There exists a variety of approaches to many versions of the problem summarized in reasonable detail in the recent survey [2]. In this section, we review two main classes of techniques which utilize classs dependent compensation/adaptation of speech feature vectors.

Two classes of techniques identify themselves by differing requirements in terms of data and computation. The Codeword Dependent Cepstral Normalization (CDCN) [1] and the Baum-Welch Codebook Adaptation (BWCA)[6] do not require a priori knowledge about the testing environment. CDCN relies on maximum likelihood estimation via EM algorithm of a model of degradation and BWCA also uses an EM-type iteration. They amount to iterative retraining of a reduced parameter set of the recognition system, therefore the computational costs are high and the performance suffers at low SNR's.

The second class of techniques, the Fixed CDCN (FCDCN) [5] and Dual Channel Codebook Adaptation (DCCA)[6] are data-driven and computationally efficient but require simultaneously recorded speech from the training and the testing environments. In many practical applications, such stereo recorded data are simply not available. These techniques have been extended to the case of unknown environments by using a basis of environments for which stereo recordings are available and codebooks can be estimated; as in the Multiple FCDCN[6], and Schwartz [8].

0-7803-3192-3/96 $5.00©1996 IEEE

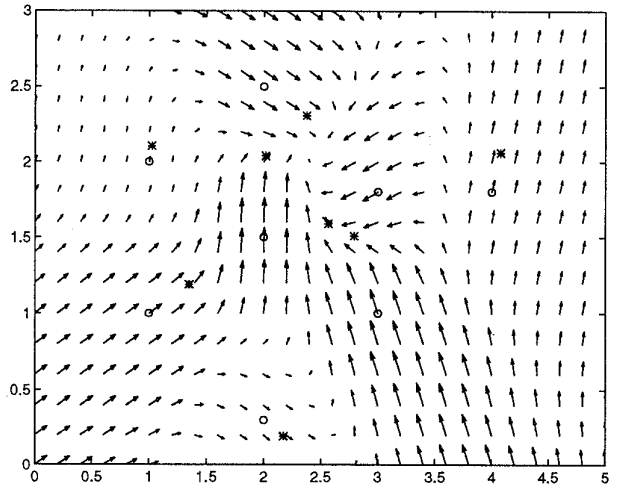

Figure 1. The difference vector field between two acoustical environments (∘ and ∗)in a 2D feature space.

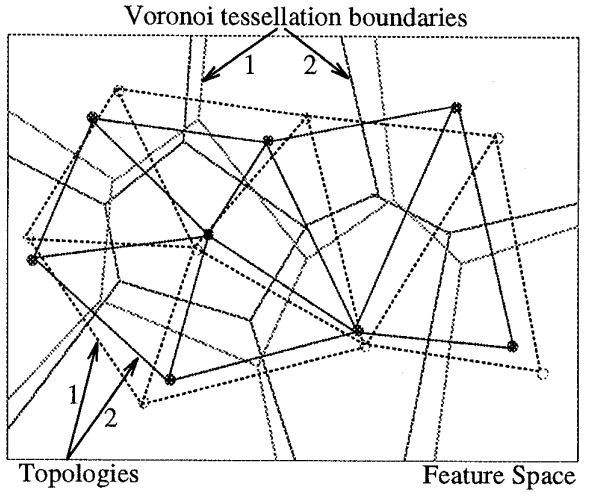

Figure 2. VQ codebooks of two environments (1 and 2) with the topology preserved.

This very important constraint has allowed both techniques to be used only with different types of microphones for which stereo recordings made in a studio environment are available. It would not be possible to collect simultaneously recorded data for a variety of environments of practical importance, such as cellular phones in moving vehicles, etc. The technique we present bypasses the need for simultaneous stereo recordings by adapting the reference VQ codebook to secondary environments while maintaining the integrity of classes by preserving the topology of the reference environment.

A second major difference is that the codebooks in the referenced approaches are fixed throughout recognition, and once an environment in the available set of environments is selected, compensation vectors or adaptation codebooks are not changed. In the technique presented here, the codebook selected among the available environment codebooks is adapted to the test environment on-line in a robust manner and therefore even if the initial match between the environments is not as good, it gets better as the codebooks get updated.

3. DISTORTION IN THE FEATURE SPACE

In the front-end of the HMM speech recognizer used in this work, a broad range of features such as frame energy, voicing, spectra and their derivatives are concatenated to form a high dimensional feature vector. Principal component analysis is applied to this high dimensional vector space to reduce dimensionality by selecting a subset of axes along which statistical variation is maximal. We denote the resulting principal component vector space by $\mathcal{F}$. Vector quantization is applied to $\mathcal{F}$, therefore, members in a class are related not only in terms of their spectra as in many other approaches, but by both static and dynamic features which determine in a more complete manner the way they are affected by the environment.

A Gaussian mixture is a common assumption for the distribution of the feature vectors of the reference environment[1, 7]. Figure 1 depicts the difference vector field between two Gaussian mixtures in a 2D feature space. Even with an acoustical degradation model as simple as linear filtering and additive Gaussian noise, the distribution of a Gaussian class of log-spectral feature vectors after degradation is no longer Gaussian [7]. Non-parametric approaches such as VQ may therefore be more reasonable choices for modeling the distribution of feature vectors in various environments. VQ-class dependent techniques model the difference vector field as the difference between the VQ codebooks for the two environments. Let us denote the codebooks as $\mathbf{X}^h = \{\mathbf{x}_k^h \in \mathcal{F} = \Re^2, k = 1, \ldots, K\}$, $h = 1, 2$. Then, a discrete approximation to the difference vector field is the set of vectors $\{\mathbf{x}_k^2 - \mathbf{x}_k^1 \in \mathcal{F} = \Re^2, k = 1, \ldots, K\}$. In this work, we are mainly interested in the estimation of these vectors in the absence of simultaneously recorded (labeled) speech. The problem is to preserve the class pairs between the two codebooks with unlabeled data.

4. TOPOLOGICALLY CONSTRAINED ADAPTIVE VQ

An acoustical environment is described by a VQ codebook, $\mathbf{X}^h = \{\mathbf{x}_k^h \in \mathcal{F}, k = 1, \ldots, K\}$ where each codevector $\mathbf{x}_k^h$ in the feature space $\mathcal{F}$ represents a class of feature vectors. The VQ codebook for the reference environment, $\mathbf{X}^{ref} = \{\mathbf{x}_k^{ref} \in \mathcal{F}, k = 1, \ldots, K\}$ is designed using the Generalized Lloyd algorithm [4]. The design criterion is the minimization of the distortion

$$D = E[d(\mathbf{x}, \mathbf{x}_w^{ref})] \tag{1}$$

where the "winner", w, is given by

$$w = \arg\min_j |\mathbf{x} - \mathbf{x}_j^{ref}|^2 \tag{2}$$

In the VQ codebooks for the testing environments, $\mathcal{X} = \{\mathbf{X}^h, h = 1, \ldots, H\}$, $\mathbf{x}_k^{ref}$ and $\mathbf{x}_k^h$ must correspond to identical acoustical classes. With a simultaneously recorded

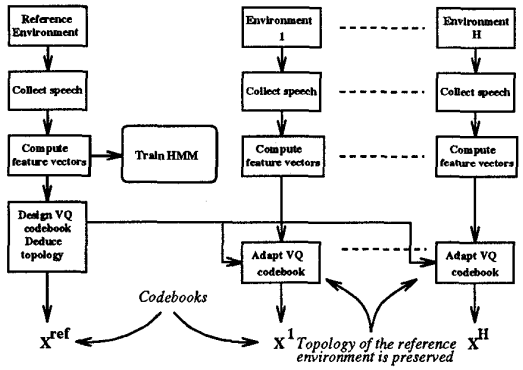

Figure 3. Off-line topology preserving VQ codebook adaptation for a set of representative environments.

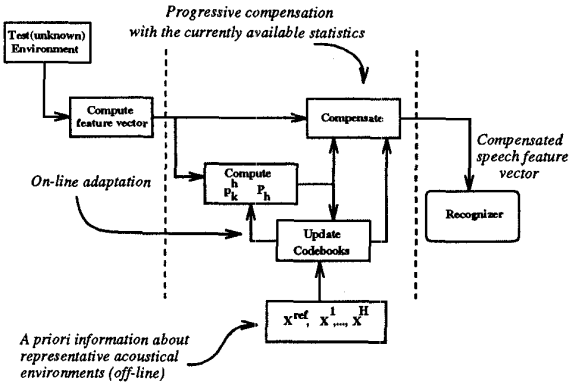

Figure 4. Compensation with both a priori information about likely environments and on-line adaptation.

stereo database, this is automatically satisfied since all the frames are labeled, and class confusion does not occur. In most practical cases, however, only unlabeled data are available.

4.1. Topology preserving codebook adaptation

To address the codebook integrity problem, we propose a codebook adaptation technique which preserves the neighborhood relations in the reference codebook by introducing a modified distortion measure for VQ. The idea of topology preservation is due to Kohonen [3]. It is one of the mathematical frameworks for the Self Organizing Map (SOM), a VQ algorithm. The topology of SOM is on a non-linear projection of the signal space onto a 1D or 2D "map", and is arbitrarily decided a priori. In our technique, the topology is directly on the feature space and simply determined by the distortion measure in the reference environment.

The topology of the reference environment, i.e. the local neighborhood relations, is captured with the neighborhood function

$$n_{ij} = \exp\left(-\frac{|\mathbf{x}_i^{ref} - \mathbf{x}_j^{ref}|^2}{2\sigma^2}\right) \forall i,j = 1,\ldots,K. \quad (3)$$

Practically, only N-closest neighbors (N=5-10) are kept and the rest of the n_{ij}'s are set to zero leading to the representation of the topology as an elastic mesh as in Figure 2. The codebook adaptation amounts to stretching of the mesh to better fit the new environment while keeping the neighborhood relations intact.

Expressed in terms of distortion, the topologically constrained VQ is the minimization of the modified distortion

$$D' = E[d(\mathbf{x}, \mathbf{x}_w^h)] + \sum_{i \neq w} n_{wi} E[d(\mathbf{x}, \mathbf{x}_i^h)] \quad (4)$$

where the "winner" w is, similarly, $w = \arg\min_j |\mathbf{x}(t) - \mathbf{x}_j^h(t)|^2$

The minimization of D' is accomplished by the Robbins-Munro stochastic approximation technique, which in the case of squared error distortion reduces to the incremental adaptation

$$\mathbf{x}_k^h(t+1) = \mathbf{x}_k^h(t) + n_{wk}(t)[\mathbf{x}(t) - \mathbf{x}_k^h(t)] \quad (5)$$

where $\mathbf{x}(t)$, $t = 0,\ldots,T$ are the available (unlabeled) data from the testing environment and the codebooks for the test environments are initialized with the codebook of the reference environment

$$\mathbf{x}_k^h(0) = \mathbf{x}_k^{ref}, \quad k = 1,\ldots,K. \quad (6)$$

Notice that in the adaptation, the neighborhood function is also dynamic in the following form

$$n_{ij}(t) = \alpha(t) \exp\left(-\frac{|\mathbf{x}_i^{ref} - \mathbf{x}_j^{ref}|^2}{2\sigma^2(t)}\right) \forall i,j = 1,\ldots,K. \quad (7)$$

The functions $\alpha(t)$ and $\sigma^2(t)$ are monotonically decreasing. For initial large values of $\sigma^2(t)$, all codevectors are updated similarly, thus the algorithm in the beginning may be regarded as an incremental version of simple mean normalization. The scale of adaptation is made finer by decreasing $\sigma^2(t)$ as increasingly more data are used.

4.2. Computation of the compensation vectors

The off-line codebook adaptation described in the previous section is carried out for a set of representative environments as shown in Figure 3. Once the codebooks are available, they are simply used as a basis in which to express the data from an unknown environment. For a discrete density HMM, the technique may be regarded as codebook adaptation, for a continuous density HMM, such as the one used in this work, it is necessary to put it in the form of a compensation algorithm as follows. Let the incoming speech feature vector(t-th frame of the utterance) from the unknown test environment be denoted as $\mathbf{x}(t)$. Then, the compensated feature vector, $\hat{\mathbf{x}}(t)$ is computed as

$$\hat{\mathbf{x}}(t) = \mathbf{x}(t) + \sum_h P_h \sum_k p_k^h(t)[\mathbf{x}_k^{ref} - \mathbf{x}_k^h(t)] \quad (8)$$

where the probability that the t-th frame belongs to Voronoi region k in the codebook h, $p_k^h(t)$, and the probability that the utterance belongs to environment h, P_h are estimated as

$$p_k^h(t) = \frac{e^{-\beta(\mathbf{x}_k^h(t) - \mathbf{x}(t))^2}}{\sum_k e^{-\beta(\mathbf{x}_k^h(t) - \mathbf{x}(t))^2}} \quad (9)$$

$$P_h = \frac{e^{-\alpha \sum_n (\mathbf{x}_w^h(t) - \mathbf{x}(t))^2}}{\sum_h e^{-\alpha \sum_n (\mathbf{x}_w^h(t) - \mathbf{x}(t))^2}}. \quad (10)$$

4.3. On-line adaptation

The initial codebook selection is a fast adaptation using a priori knowledge about likely representative environments. A new testing environment may not always fit the available codebooks to give a satisfactory performance. In such cases, on-line adaptation to the new environment may be accomplished by utilizing the testing environment's data during compensation via the same stochastic approximation. This is shown in the block diagram of the compensation in Figure 4. In this way, even if the initial match between the environments is not as good, it gets better as the codebooks get updated.

5. EXPERIMENTAL RESULTS

Results are presented on continuous digit recognition and voice dialing in the McCaw Cellular Corpus. The corpus consists of data collected over cellular channels by using two types of microphones: a hand-held, close talking microphone and a hands-free, visor mounted microphone together with land-line collected speech data. The land-line and hand-held microphone parts of the corpus are mostly clean telephone speech comparable in quality to VAA corpus. The hands-free microphone part of the corpus, however, is significantly noisier than the rest.

environment	*no. of* utt.'s	*error baseline*	*error w/ MN*	*error w/ AVQ*
VAA2	1390	4.1	4.1	4.2
land line	282	4.5	4.4	4.7
hand held	283	6.0	6.0	6.1
hands free	246	23.8	17.8	13.6

Table 1. Results of the speaker independent digit recognition experiment.

environment	*no. of* utt.'s	*error baseline*	*error w/ MN*	*error w/ AVQ*
land line	696	3.4	3.4	3.7
hand held	688	4.7	4.8	5.4
hands free	650	16.5	13.4	10.6

Table 2. Results of the speaker dependent voice calling experiment.

The first experiment investigates the effectiveness of the compensation algorithm in normalizing the McCaw speaker independent digit recognition data to improve recognition using models trained on the VAA1 corpus. The codebook size for which the results are reported here is 16. The codebooks were trained on data sets in the McCaw and VAA corpora disjoint from the model training and testing sets for which the recognition results were obtained. The results in Table 1 indicate that the normalization does not disturb the reference environment (VAA) appreciably, nor the land line and hand held environments which are close to the VAA. There is a 43% decrease in the error of the hands free microphone.

A similar experiment was carried out on the speaker dependent portion of the McCaw database. Table 2 summarizes the average results for 30 speakers each uttering 10 names in a voice calling application in which the land-line is the reference environment. The reference and clean environments are again not disturbed appreciably and there is a 36% decrease in the error of the hands free microphone.

6. CONCLUSION

We introduce a topological constraint to VQ design which allows adaptation to environments while maintaining the integrity of the class memberships without using simultaneously recorded speech. This adaptation is used both off-line to form a basis of representative environments and on-line to further improve the initial match between the codebook and the testing environment. The technique is useful in applications where stereo recorded data are not available and the computational loads must be kept low. It decreases the word error for continuous ten digit recognition of cellular hands free microphone speech with land line trained models from 23.8% to 13.6% and the speaker dependent voice calling sentence error from 16.5% to 10.6% in the McCaw cellular corpus.

REFERENCES

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.

[2] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication,* **16**, 1995 pp. 261-291.

[3] T. Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, Springer-Verlag, Berlin 1995

[4] Y. Linde, A. Buzo, R.M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, January 1980.

[5] F.H. Liu, R.H. Stern, A. Acero, P.J. Moreno, "Environment Normalization for Robust Speech Recognition using Direct Cepstral Comparison," *ICASSP-94*, pp. 61-64, April 1994

[6] F.H. Liu, "Environmental Adaptation for Robust Speech Recognition," Ph. D. Thesis, ECE Department, CMU, July 1994

[7] P.J. Moreno, B. Raj, E. Gouvêa, R.M. Stern, "Multivariate Gaussian Based Cepstral Normalization", *ICASSP-95*

[8] R. Schwartz, T. Anastakos, F. Kubala, J. Makhoul, L. Nguyen, G. Zavaliagkos, "Comparative Experiments on Large Vocabulary Speech Recognition," *Proc. ARPA Human Language Technology Workshop,* Plainsboro, New Jersey, March 1993.