CLASSIFYING USER ENVIRONMENT FOR MOBILE APPLICATIONS USING LINEAR AUTOENCODING OF AMBIENT AUDIO

Robert G. Malkin and Alex Waibel

Interactive Systems Laboratories, Carnegie Mellon University
{malkin,ahw}@cs.cmu.edu

ABSTRACT

Many mobile devices and applications can act in context-sensitive ways, but rely on explicit human action for context awareness. It would be preferable if our devices were able to attain context awareness without human intervention. One important aspect of user context is environment. We present a novel method for classifying environment types based on acoustic signals. This method makes use of linear autoencoding neural networks, and is motivated by the observation that biological coding systems seem to be heavily influenced by the statistics of their environments. We show that the autoencoder method achieved a lower error rate than a standard gaussian mixture model on a representative sample task, and that a linear combination of autoencoders and GMMs yielded better performance than either alone.

1. INTRODUCTION

Many mobile devices and applications can act in context-sensitive ways, but rely on explicit human action for context awareness. It would be preferable if our devices were able to attain context awareness without human intervention. Thus, we are interested in developing systems which can analyze perceptual data to draw inferences about user context. One important aspect of user context is environment. A cellphone that is aware of its user's environment, like the Connector system under development in the CHIL project [1], would be able to switch between notification modes automatically; for example, switching to silent mode if the user is in a restaurant or theater, or halting notification entirely if the user is attending a meeting or lecture.

To perform environment classification, we choose to focus on the audio signal. We make this choice for several reasons. The audio signal is relatively low-bandwidth and has modest processing and storage requirements, and good-quality audio sensors are cheap and robust. More importantly, the audio signal is largely unaffected by potentially spurious changes in aspect, position, or lighting, and the environment almost always leaves strong evidence in the audio signal — often strong enough that humans talking on the telephone can determine whether the caller is in an office, car, city street, or airport.

We present a novel method for classifying environment types based on acoustic signals. This method makes use of linear autoencoding neural networks, and is motivated by the observation that biological coding systems seem to be heavily influenced by the statistics of their environments. We show that the autoencoder method outperforms a standard gaussian mixture model (GMM) on a representative sample task, and that a linear combination of autoencoders and GMMs yields better performance than either alone.

The remainder of this paper is organized as follows. Related work is discussed in Section 2, the linear autoencoder model is presented in Section 3, and the evaluation procedure is presented in Section 4. Discussion follows in Section 5.

2. RELATED WORK

Clarkson and Pentland studied user context awareness using audio, video, and other sensory streams [2] [3], [4], [5], [6] in the context of a system designed to extract personal life patterns from sensory data. This system employed feature-level fusion and HMM clustering techniques to learn common scenarios in everyday life.

Ellis and Lee presented a personal archiving system [7] that used an unsupervised spectral clustering technique to analyze environmental audio. Their system achieved 61% accuracy on a task similar to the one described here.

3. LINEAR AUTOENCODER NETWORKS

As noted above, we were motivated to employ a linear autoencoder for this task by the observation that biological perceptual systems seem to display filter structures that are similar to those derived from information-theoretic procedures for optimal coding [8], [9], [10], [11] which seek to maximize statistical independence of filter components. This fact suggested that a model based on optimal coding

This research is supported by the European Commission CHIL project (http://chil.server.de) under contract No. 506909.

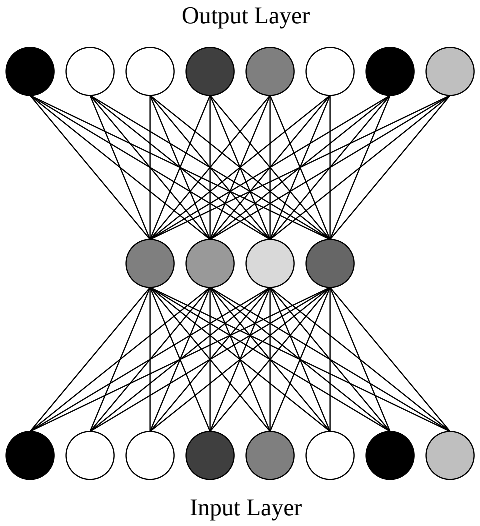


Fig. 1. A Linear Autoencoder

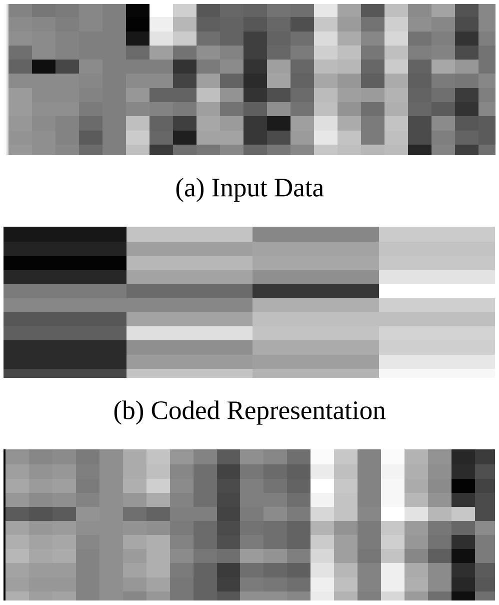
might be able to perform environment classification. That is, an optimal coder exposed only to sounds from one specific environment would display structures characteristic to that environment. These coders should be able to code and reconstruct signals from their own environments quite well, but should fare poorly on signals from other environments. If this hypothesis is correct, measuring the signal reconstruction error over a bank of environment-specific optimal coders could be a reasonable procedure for acoustic environment classification.

A simple model capable of performing optimal coding under an assumption of data gaussianity is a linear autoencoding neural network, or autoencoder [12] [13], an example of which is shown in Figure 1. The autoencoder is a standard feed-forward neural network with a linear transfer function trained on the identity function; that is, the desired output is the same as the input. An example of several frames of data passed through a network trained in this way is shown in Figure 2. In this example, the network reproduced the gross features of high-dimensional data using only four hidden units. It has been shown that training such a network using a sum-of-squares error function leads the weights in the hidden layer to approximate the subspace spanned by the N largest principal components of the data [13]. This implies that the hidden unit activations are decorrelated, which under the gaussian assumption is equivalent to statistical independence.

4. EVALUATION

4.1. Data Collection

We evaluated the autoencoding method on a corpus of audio recordings containing eleven different environments. The environments were chosen to represent a variety of indoor and outdoor types. Data collection personnel recorded ten-minute scenes in these environments, with a variety of different locales. The environments we chose to study were *apartment* (5 scenes), *office corridor* (1), *elevator* (2), *lecture* (5), *meeting room* (3), *office* (4), *outdoor* (7), *outdoor raining* (6), *restaurant* (2), *theater* (2), and *vehicle* (8). The outdoor class included mainly areas with little or no street-type noise, and the vehicle class included cars, buses, and trains. The data collectors were instructed to keep foreground speech to a minimum so that the ambient sound textures would be the predominant sources of acoustic energy.



(a) Input Data

(b) Coded Representation

(c) Autoencoder Output

Fig. 2. Linear Autoencoder Example

Data were collected using a small digital voice recorder and a Sony ECM-717 stereo microphone. The data were resampled at 16 kHz, 16-bit mono for use in this experiment. To help ensure the predominance of ambient sound in the training and test sets, we calculated the mean power of each recording and selected for further study only those segments that were quieter than average (this amounts to the inverse of the event detection procedure described in [2]). After applying this procedure, the data were divided into a training set and a test set; 80% of the segments were selected for training; the remainder for testing. The number of segments and time per class in the training and test sets are shown in Figure 3. There were a total of 7299 segments in the training set, amounting to 218 minutes. The test set contained 1814 segments, amounting to 49 minutes.

Class	*Training Set*		*Test Set*	
	Segs	*Total Time*	*Segs*	*Total Time*
apt	727	1563.47s	180	414.23s
cor	116	115.93s	28	31.43s
elv	349	558.97s	86	103.47s
lec	1242	1676.06s	310	422.81s
mtg	884	1286.13s	220	340.12s
ofc	661	1573.41s	164	327.69s
out	1076	2527.63s	268	350.64s
rng	634	1664.88s	158	387.85s
rst	253	205.43s	62	54.25s
tht	400	596.58s	100	169.07s
veh	957	1356.73s	238	358.31s

Fig. 3. Data Set Details

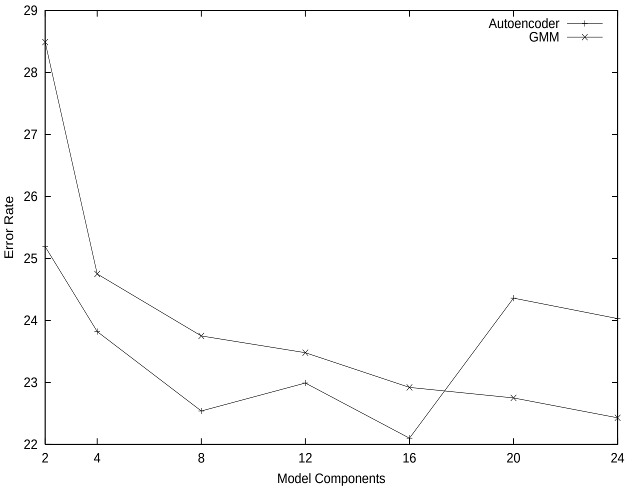


Fig. 4. Test Set Error Curve

4.2. Feature Extraction

From the signal, we extracted 64 MFCCs, plus the spectral centroid, at a rate of 100 frames per second. The spectral centroid is computed as $c = \sum_i f_i a_i / \sum_i a_i$ (where f_i is the i^{th} frequency and a_i is the amplitude at f_i) and is intended to serve as an estimate of the perceived "brightness" of a sound. Both MFCCs and brightness were normalized to zero mean and unity variance over the entire training set, and were combined into a single feature vector. This vector was then compressed to 35 dimensions using PCA, which preserved 75% of the variance. Finally, the PCA features were each scaled to unity variance, producing a sphered dataset.

4.3. Training

We trained linear autoencoders and GMMs for each class. We used seven different training configurations to test the effect of varying numbers of parameters. Specifically, we trained models with 2, 4, 8, 12, 16, 20, and 24 hidden units or gaussians. Each autoencoder was initialized with random weights between -0.05 and 0.05 and trained on the whitened features for 100 iterations using a batch-mode backpropagation algorithm with an adaptive learning rate initialized at 0.05, a momentum term of 0.045, and batch shuffling. Each GMM was trained on the same features for 20 iterations using the neural gas algorithm [14] a soft-clustering variant of the k-means algorithm, with a starting temperature of 0.5 and a cooling rate of 0.01.

4.4. Results

For each test segment, we calculated both autoencoder and GMM scores. In each case, the hypothesis was taken to be the class of the model producing the optimum score. The error curves for this experiment are shown in Figure 4. The 16-unit autoencoder achieved a 22.1% error rate, while the 24-gaussian GMM achieved a 22.43% error rate. In general, the autoencoders outperformed GMMs with the same number of parameters for models with fewer than 20 components.

After running this experiment, we examined the confusion matrices of the best autoencoders and GMMs. We discovered that there were some systematic errors that were unique to one classifier or the other. The per-class precision and recall is shown in Figure 5. Several large differences present themselves; the restaurant recall and precision and the corridor and theater precision are notable.

Given this mismatch, we postulated that a hybrid model using both autoencoders and GMMs might outperform either model alone. To test this hypothesis, we carried out another test in which we simply added weighted autoencoder and GMM scores together. The results of this test are shown in Figure 6, with the 2- and 4-component models omitted. The lowest error for the hybrid system shown was 19.95%, achieved with 16 hidden units/gaussians and a score weighting of 3-2 in favor of the autoencoders. A trivially better performance of 19.78% was achieved by combining 16-unit autoencoders with 24-gaussian GMMs with equal score weighting. The best hybrid model thus performed 2.32% absolute better than the best autoencoder model, and 2.64% absolute better than the best GMM.

We conducted no studies on this dataset to determine how well humans perform at this task. However, in a smaller pilot study consisting of only one recording each from six different classes (office, cafeteria, car, lecture, city street, CMU campus), we found that untrained human listeners achieved error rates between 20% and 30%.

Class	*Autoencoder*		*GMM*	
	Precision	Recall	Precision	Recall
apt	87.97	89.44	81.00	90.00
cor	**74.19**	82.14	**92.59**	89.28
elv	68.90	95.34	65.34	96.51
lec	80.88	77.45	77.88	78.38
mtg	86.40	80.90	96.56	75.45
ofc	93.86	93.29	96.27	94.51
out	78.01	55.59	86.14	53.35
rng	69.69	87.34	75.27	86.70
rst	**43.47**	**64.51**	**32.58**	**93.54**
tht	**71.99**	90.00	**93.40**	84.99
veh	76.05	68.06	75.37	63.02

Fig. 5. Precision and Recall per Class

GMM Weight	*Hidden Units / Number of Gaussians*				
	8	12	16	20	24
0.1	22.15	22.26	21.60	23.31	22.65
0.2	21.82	21.71	21.11	22.54	21.88
0.3	21.38	21.00	20.00	22.04	21.49
0.4	21.05	20.50	**19.95**	21.38	21.49
0.5	20.72	20.39	20.22	21.27	21.33
0.6	20.66	20.39	20.28	21.16	21.38
0.7	21.00	21.22	20.61	21.38	21.44
0.8	21.38	21.60	21.33	21.60	21.99
0.9	22.81	22.59	21.88	22.10	22.26

Fig. 6. Error Rates for Autoencoder/GMM Hybrid

5. DISCUSSION

We demonstrated that linear autoencoder networks outperformed GMMs on an environment classification task, validating the hypothesis that procedures based on optimal coding are effective for this task. Further, we demonstrated that the errors made by autoencoder networks and GMMs were sufficiently different that a simple linear combination of the two models yielded reduced error rates.

In future work, we will attempt to improve system accuracy, coverage, and utility. We will build a corpus better suited to the CHIL Connector task, and include a wider variety of specific locales from the US and Europe. Further, we intend to make a more systematic study of human performance on the environment classification task in order to provide a gold standard against which to measure machine performance.

6. REFERENCES

[1] A. Waibel, H. Steusloff, R. Stiefelhagen, and the CHIL Project Consortium, “CHIL: Computers in the human interaction loop,” in *International Workshop on Image Analysis for Multimedia Interactive Services*, 2004.

[2] B. Clarkson and A. Pentland, “Extracting context from environmental audio,” in *Proceedings of the 2nd International Symposium on Wearable Computers*, 1998.

[3] B. Clarkson et. al., “Auditory context awareness via wearable computing,” in *Proceedings of the Perceptual User Interfaces Workshop*, 1998.

[4] B. Clarkson and A. Pentland, “Unsupervised clustering of ambulatory audio and video,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[5] B. Clarkson and A. Pentland, “Framing through peripheral perception,” in *Proceedings of the International Conference on Image Processing*, 2000.

[6] B. Clarkson, *Life Patterns: Structure from Wearable Sensors*, Ph.D. thesis, MIT, 2002.

[7] D. Ellis and K.S. Lee, “Minimal-impact audio-based personal archives,” in *First ACM Workshop on Continuous Archiving and Recording of Personal Experiences*, 2004.

[8] M.S. Lewicki, “Efficient coding of natural sounds,” *Nature Neuroscience*, 2002.

[9] A.J. Bell and T.J. Sejnowksi, “The ’indepdendent components’ of natural scenes are edge filters,” *Vision Research*, 1997.

[10] H.B. Barlow, “Possible principles underlying the transformation of sensory messages,” in *Sensory Communication*, W.A. Rosenbluth, Ed. MIT Press, 1961.

[11] J.J. Atick, “Could information theory provide an ecological theory of sensory processing?,” *Network: Computation in Neural Systems*, 1992.

[12] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley and Sons, 2001.

[13] J. Hertz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation*, Perseus Books Publishing, 1991.

[14] T.M. Martinez and K.J. Schulten, *Artificial Neural Networks*, chapter A ”Neural-Gas” Network Learns Topologies, North-Holland, 1991.