Proc APSIPA 2012
Hollywood, CA

Composite-DBN for Recognition of Environmental Contexts

Selina Chu*, Shrikanth Narayanan† and C.-C. Jay Kuo†
* Oregon State University
E-mail: selina@eecs.oregonstate.edu
† University of Southern California
E-mail: {shri, cckuo}@sipi.usc.edu

***Abstract*—People's behaviors are usually dictated by their surroundings. The surrounding environment affects the character and disposition of the people within it. The goal of our work is to automatically recognize the type of environments one is in. In this paper, we introduce a hierarchical structure to recognize environments using the surrounding audio. We can use this structure to discover high-level representations for different acoustic environments in a data-driven fashion. Being able to perform such function would allow us to better understand how we could utilize such information to assist in predicting a person's emotion or behavior. To accurately make an informative decision about behaviors or emotions, it is important to have the ability to differentiate between different types of environments. Environmental sound contains large variances even within a single environment and is constantly changing. These changes and events are dynamic and inconsistent. The goal is to come up with models that is robust enough to generalize to different situations. Learning a hierarchy of sound types would improve and clarify problems caused by the confusion between multiple acoustic environments with similar characteristics. We propose a framework for a composite of deep belief networks (composite-DBNs) as a way to represent various levels of representations and to recognize twelve different types of common everyday environments. Experimental results demonstrate promising performance in improving the state of art recognition for acoustic environments.**

I. Introduction

There has been recent interest in behavioral informatics, such as recognizing emotions using acoustic features . For example, Polzehl et al. exploit both linguistic and acoustic feature for modeling anger recognition in speech [1] and Busso et al. uses facial expression and speech to recognize emotions [2]. People's general behavior is typically dictated by their surrounding environments. Therefore, to accurately formulate an informative prediction about a behavior or emotion, it is necessary to determine the type of situation that person is in. For example, if someone is in a business meeting or in a class room, typical behaviors will be more subdued as compared to being at the beach or in an amusement park. Our goal is to investigate the utility of the environmental surroundings and their influence on people's behavior. The first step toward this goal is having the ability to determine the type of environment one is situated in.

The nature of environment is dynamic and constantly changing. The difficulty comes as natural environments contain large variances, even within a single environment, making it difficult to to build models for. Despite these differences, humans can distinguish and contextualize them. For example, humans could easily differentiate between sounds originating from outside on the streets, inside a restaurant, coffee shop, or train station, etc. We could mostly agree that it is relatively easy for us to identify something as a restaurant environment even when presented with audio recordings of varying restaurant locations or settings. This implies that there are commonalities between different locations of the same situation or context. This work attempts to uncover these commonalities despite the noise and variance that comes with the changing environment so that we are able to utilize the surrounding information in a more tangible manner. If we could learn some commonalities between certain related environments, then we could use this information, fusing with other features (e.g. speech or video), to make a more informative decision by adding another level of confidence.

The advantage of using audio for recognition is that it is computationally cheaper to process as compared to video, for example, being more practical for real-time applications. In addition, video recordings requires much world knowledge and their quality is dependent on the lighting and angle of the video camera. However, realistic environmental sound is typically noisy and contains similar characteristics between different environments. Another challenge is that some types of sounds are perceptually different but the statistical features are similar (or vice versa). For example, ocean wave and car crash sound is perceptually different sounding, but the temporal and spectral features are similar. Two different types of motors might sound very similar, but the extracted features are different [3].

In this work, we investigate the use of a richer generative model based method for acoustic environmental classification and to discover high-level representations for different acoustic environments in a data-driven fashion. Specifically, we consider a composite of deep belief networks (DBNs) as a way to model environmental audio and investigate its applicability with noisy auditory data for robustness and generalization.

II. Classification of acoustic environment

Natural environmental sounds, in addition to being noisy and having large amount of variance, have no divisible or clear structure between different types. These audio requires representation by complex models, but traditional audio classifiers still deteriorate dramatically when using realistic environmental sound that are noisy or have overlapping classes. Traditional classification methods, like Gaussian mixture models (GMM) [4] and Hidden Markov Models (HMM) [5], are commonly used classifiers for audio classification in general. HMMs have been extensively used in speech and works well with sounds that change in duration. Since environmental sounds or general ambient sounds lack such temporal structure or phonetic structure that speech has, there is no set alphabet that allows for slices of non-speech sound to be divided into, making HMM-based methods difficult to implement. Non-linear classifier like SVM and traditional neural network have shown to work well for audio classification to discriminate on non-linear separable classes [6]. However, they do not scale well to long feature vectors as input and larger number of classes (e.g. over 10 classes). Since they are not as efficient as GMMs or HMMs, these non-linear classifiers have been utilized to a lesser extent.

A. Deep Belief Networks (DBN)

In recent years, there has been a large interest in deep learning and using neural network with recent introduction of a fast greedy layer-wise unsupervised learning algorithm by Hinton et al [7]. DBNs have been applied to music audio [8] and to learn features for speech recognition [9]. The idea of using this method is to learn some abstract representations of the input data in a layer-wise fashion using unsupervised learning, which then can be used as input for supervised learning in tasks such as classification and regression.

DBN is a neural network constructed from multiple layers of Restricted Boltzmann Machines (RBMs). A RBM is a bipartite graph composed of a layer of stochastic visible units v and a layer of stochastic hidden unit h. Many RBMs can be stacked on top of each other by linking the hidden layer of one RBM to the visible layer of the next RBM, forming a multilayer neural network. Previously, traditional neural network were trained using gradient descent. Using gradient descent, however, makes neural networks difficult or impossible to train. Hinton proposes a greedy layer-wise unsupervised pre-training phase, which in [7, 10] showed that this unsupervised pre-training builds a representation from which made it possible to perform supervised learning by fine-tuning the resulting weights using gradient descent learning (similar to traditional neural network learning). In other words, the unsupervised stage sets the weights of the network to be closer to a good solution than random initialization, thus avoiding local minima which made occur when using supervised gradient descent.

The DBN is trained in two phases. The pre-training phase considers each layer (an RBM) separately and trains layers closest to the input layer first. It takes the output of the first layer and use it as input to the next layer, and so forth. It uses the greedy layer-wise Contrastive Divergence (CD) pre-training for initializing weights. The overall pre-training process is repeated several times, layer by layer, obtaining a hierarchical model in which each layer captures strong high-order correlations between its input units. This phase allows the DBN to make use of unlabeled data in an unsupervised manner. The second phase of training is a supervised, global fine-tuning phase that is similar to training traditional neural network training. Gradient descent is used to obtain a fine-tuning of the parameters for optimal reconstruction of the input data. For more detailed treatment of DBN, we refer readers to [7].

III. DBN for Environmental Sound

A. Experimental Setup

Environment types were chosen so that they are made up of ambient sounds of a particular environment, composed of many sound events. We do not consider each constituent sound event individually, but as many properties of each environment. We used recordings of natural (unsynthesized) sound clips obtained from [11]. Our auditory environment types were chosen so that they are made up mostly of non-speech and non-music sounds. One could think of it as background noise of a particular environment, composed of many sound events. The twelve environment types considered were: 1) *Inside casino*, 2) *Playground*, 3) *Nature-nighttime*, 4) *Nature-daytime*, 5) *Inside restaurants*, 6) *Next to rivers/streams*, 7) *Train passing*, 8) *Inside vehicles*, 9) *Raining*, 10) *Street with traffic*, 11) *Ocean waves*, and 12) *Thundering*.

The sound clips used are of varying lengths (1-3 minutes long) and are later processed by dividing them up into 3-second segments and downsampled to 22050 Hz sampling rate, mono-channel and 16 bits per sample. Each 3-sec segment makes up an instance for training/testing. The audio was analyzed using a rectangular window of 512 points (23.3 msec) with 50% overlap. We represented the audio using Mel-frequency cepstrum coefficient analysis (MFCC) [5] and Matching Pursuit (MP)-features [12]. MFCC is the most common feature representation for audio and has shown to work relatively well for speech and music, but their performance degrades in the presence of noise. *MP-features*, which is a feature extraction method specifically proposed for environmental sounds [12]. Previous research on audio features have shown that using MP-features for environmental sounds proved to be successful in aiding with classification of unstructured environmental sound. Since MP-features are discrete values, we discretize MFCCs using equal frequency discretization method from [13], which resulted with 174-dimensional feature vector for each data sample. We kept samples originated from the same source separate from one another. With this setup, none of the training and test items originated from the same source. Since the recordings were taken from a wide variety of locations, the ambient sound might have a very high variance. The only preprocessing we performed on the data was verifying that they were not saturated and to remove the silent parts from beginning and end of the files. The data used consisted of two different sets: *Set A*: Samples were selected so that they are more homogeneous within each type of environment. The samples

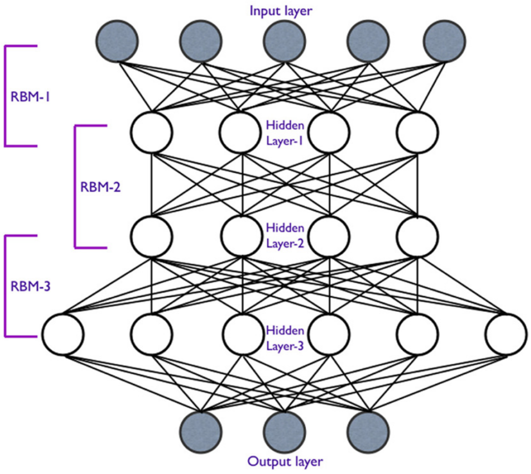

Fig. 1. Configuration of DBN used

are also enforced so that each type of sound tend to be distinctively sounding different from one another, which minimized overlaps as much as possible. Each environment contained at least four separate source recordings, and segments from the same source file were considered a batch. We use three batches for training and one batch for testing, which leads us to perform a 4-fold cross validation for the features. There are around 10-15 files for each environment. *Set B*: Samples that are *related* to the same environment types as in *Set A*, but more diverse sounds of the same class . There are no restrictions on the data, like homogeneity within each class or minimal overlap between types.

We use a DBN with three hidden layers with 100 hidden units for the first and second layers and 450 hidden units for the third layer. The input layer consists of 174 units which corresponds to the dimension of the feature vector and 12 output units for the number of target classes. We use a learning rate of 0.1. A schematic representation is given in Fig. 1.

B. Classification

To analyze the performance of DBN for environmental sounds, we experimented with various settings and compared the results to those obtained by using GMM. For GMM classifiers, we use 5 mixtures throughout all of our experiments. Note that it is possible to tailor the number of mixtures to each class of data. The improvement has shown to be negligible but will cause the classifier to be specialized to the training data. Therefore, we decided to just use a set number of mixtures throughout all of our experiments.

We compared results on classification of: 1) using only *Set A* and 2) using both *Set A* and *B* The results are shown in Table I.

As expected, the more complex DBN model perform slightly better than GMM. We can see that when we included *Set B*, it increased the classification accuracy of the DBN, but the opposite occurred in the case with using GMM. It has difficulties handling the extra data.

TABLE I
CLASSIFICATION ACCURACIES COMPARING DBN AND GMM (IN %)

Data set used	DBN	GMM
A	67.9	64.8
A + B	79.6	41.7

When including the more diverse training data *Set B* to the DBN method, the performance improved for eight classes, but reduced the results in three classes, ranging from 5-23%. The improvement was most significant for *Near river* from 10% to 95% and for *restaurant* from 20% to 87.5%. For GMM, the performances of *casino*, *nature nighttime* and *Ocean waves* were reduced to 0%. The misclassifications seem to gravitate toward classes *Near river*, *On vehicle*, and *Raining*. The bias might be deduced from the characteristics of these three classes being somewhat more constant and homogeneous sounding, particularly for *Near river* and *Raining*. The resulting classification GMM are biased toward classifying the test data into them, as compared to models that are more diverse.

Using *A* and *B* for training provide better performance overall for using DBN. By introducing variability into the data, the performance for certain classes might slightly degrade, but is limited. The increase in the recognition ability for other classes outweighs the amount that has decreased. Even when introducing additional information that might be noisy, DBNs performance does not suffer as much and less dramatic than GMM.

IV. COMPOSITE-DBNS

Learning a hierarchy of sound types could improve and clarify problems caused by the confusion of an acoustic environment with similar characteristics. For example, a restaurant and a shopping mall, both shared characteristics of being indoors and in a crowd with people talking, but the restaurant contains clanking of utensils, whereas in a mall setting, there are footsteps and shuffling of feet. The use of suitable hierarchies would also allow us to assign confusing samples to more practical and general classes. For example, we could group *Near river* and *Raining* into fluidic type and place *Restaurant* and *Casino* into an indoor-crowd type. It allows environments to be grouped together using the heuristic that a *good* grouping of the two classes would minimize the number of misclassified.

To automatically build this hierarchy of DBNs in some optimal way, we investigate on learning audio structural models for the general environments utilizing a hierarchy of sound types. DBN actually supplies us with a simple method to accomplish this.

The idea is to utilize the combination of activations between the last hidden layer and each of the target output units of a trained DBN. Let us consider the experiment with the same set up as in Sec. III. After the DBN is trained, each target unit would be equipped with a set of activations. Using these activations as features, we calculate the pairwise distance between pairs of activation set for each environment type using cosine similarity measure. From there, we generated a hierarchical

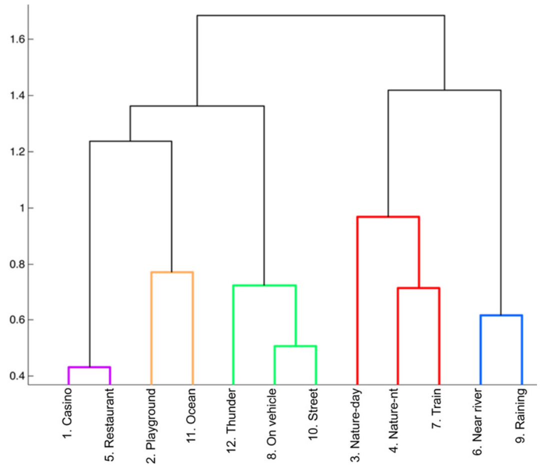

Fig. 2. Dendrogram using hierarchical clustering with cosine similarity measure, based on activations of a trained 12-class DBN (distance along the y-axis depicts their measure of similarity)

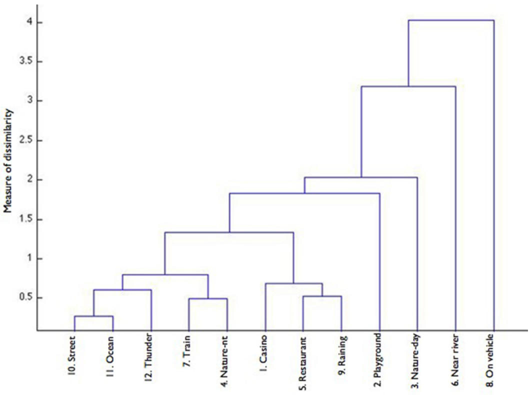

Fig. 3. Dendrogram using hierarchical clustering with cosine similarity measure using MP-features and MFCC directly (without using DBN).

clustering from the resulting pairwise distances using averaged linkage for the hierarchical structure of environment types. A dendrogram from the result is depicted in Fig 2

A composite DBN framework was built based on the results provided by the hierarchical clustering of the activations found from the 12-class classification in Sec. III-B This composite DBN yielded an average of 91.9%. To the best of our knowledge, this is a significant improvement on classification over previous proposed methods discriminating this many different types of general environmental sounds.

As a baseline for comparison, we eliminated the DBN step and obtain a hierarchical clustering founded from using only the basic audio features (MP-features and MFCCs) that was used for training the DBNs. For the distance measure, we also used cosine similarity and averaged linkage to generate the clusters. Using the average of each cluster, we obtain a dendrogram to illustrate the distances between each class, as depicted in Fig. 3. We can observe that the dendrogram created is unevenly branched, meaning that the clusters are considered to be very similar to each other, thus making it difficult to separate and permit the branches to be distributed more widely. The high level grouping of the classes are grouped together more acoustically sounding. For example *Ocean* might be closer to *Thundering* due to the crashing sound and both *Train* and *Nature-nighttime* have high frequency periodic sounds. For comparison, a composite DBN was also created based on these results Fig. 3, which yields an average classification accuracy of 51.76%.

V. Conclusions and Future Work

This paper proposes a framework for generative modeling of environmental sound using DBN in a hierarchal structure. Our framework demonstrates the ability to model different environmental sound types despite overlapping and noisy data. Experimental results demonstrate promising performance in improving the state of art recognition for audio environments. It is encouraging that we could utilize more unrestrictive data to improve generalization.

The use of additional information of a person's surrounding environment would increase the performance in multimodal emotion recognition. Learning a hierarchical structure of sound types would alleviate the confusion between multiple acoustic environments with similar characteristics. We can develop multimodal fusion models to include, for example, speech and knowledge of the environment. Utilizing additional information about the environment would allow us to make a more informative decision by adding another level of confidence in any automatic recognition process.

References

[1] T. Polzehl, A. Schmitt, F. Metze, & M. Wagner, Anger Recognition in Speech Using Acoustic and Linguistic Cues, *Speech Communication*, Vol. Special Issue: Sensing Emotion and Affect - Facing Realism in Speech Processing, 2010.

[2] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, & S. Narayanan, Analysis of emotion recognition using facial expressions, speech and multimodal information, *Proc of ICMI*, 2004.

[3] R. Cai, L. Lu, H.-J. Zhang, & L.-H. Cai, Improve audio representation by using feature structure patterns, *Proc of ICASSP*, 2004.

[4] C. Bishop, Neural networks for pattern recognition. *Oxford University Press*, 2003.

[5] L. Rabiner & B.-H. Juang, Fundamentals of speech recognition, *Prentice-Hall*, 1993.

[6] S. Chu, S. Narayanan, C.-C. J. Kuo, & M. Matarić, Where am I? scene recognition for mobile robots using audio features, *ICME*, 2006.

[7] G. Hinton, S. Osindero, Y. & -W. Teh, A fast learning algorithm for deep belief nets, *Neural Computation*, 18, 15271554, 2006.

[8] P. Hamel, S. Wood & D. Eck, Automatic identification of instrument classes in polyphonic and polyinstrument audio, *Proc of ISMIR*, 2009.

[9] H. Lee, Y. Largman, P. Pham, & A. Ng, Unsupervised feature learning for audio classification using convolutional deep belief networks, *Proc of NIPS*, 2009.

[10] Y. Bengio, P. Lamblin, D. Popovici, & H. Larochelle, Greedy layer-wise training of deep networks. *Proc of NIPS*,2007.

[11] The BBC sound effects library - original series. http://www.sound-ideas.com/bbc.html.

[12] S. Chu, S. Narayanan, and C.-C. J. Kuo, Environmental Sound Recognition With TimeFrequency Audio Features, In *IEEE Transactions on Speech, Audio, and Language Processing*, vol. 17 no. 6, pg. 1142-1158, 2009

[13] J. Dougherty, R. Kohavi, & M. Sahami, Supervised and unsupervised discretization of continuous features. *Proc of ICML*, 1995