LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-663102

A Scalable Feature Learning and Tag PredictionFramework for Natural Environment Sounds

P. Sattigeri, J. J. Thiagarajan, M. Shah, K. N. Ramamurthy, A. Spanias

October 21, 2014

Asilomar Conference on Signals, Systems, and Computers
Pacific Grove, CA, United States
November 2, 2014 through November 5, 2014

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

A Scalable Feature Learning and Tag Prediction Framework for Natural Environment Sounds

P. Sattigeri[1], J.J. Thiagarajan[2], M. Shah[1], K.N. Ramamurthy[3], and A. Spanias[1]
[1]SenSIP Center, School of ECEE, Arizona State University
[2]Lawrence Livermore National Laboratory
[3]IBM Thomas J. Watson Research Center

***Abstract*—Building feature extraction approaches that can effectively characterize natural environment sounds is challenging due to the dynamic nature. In this paper, we develop a framework for feature extraction and obtaining semantic inferences from such data. In particular, we propose a new pooling strategy for deep architectures, that can preserve the temporal dynamics in the resulting representation. By constructing an ensemble of semantic embeddings, we employ an ℓ_1-reconstruction based prediction algorithm for estimating the relevant tags. We evaluate our approach on challenging environmental sound recognition datasets, and show that the proposed features outperform traditional spectral features.**

I. Introduction

Building concise representations for multimedia is crucial for applications ranging from scene recognition, retrieval and personal life-logging systems to field robot navigation. A convenient approach to representing the content of multimedia data is to associate textual tags that can describe the underlying semantics. However, the effectiveness of semantic inference completely relies on the richness of the data, and there is an increasing need to utilize multiple data modalities to understand a physical process. For example, short time events such as an explosion can be effectively captured in the audio data, while the visual data from this event may be incomplete or unavailable due to the limited view or slow response of the camera. Consequently, research efforts have been focused on building tools that are specifically adapted for obtaining inferences from each data modality.

In this paper, we consider the use of novel machine learning tools to model and obtain semantic inferences from natural environment sounds. The growing interest in technologies for wearable computing, automatic life logging, and predictive inferences in robotics presents a huge potential for algorithms that characterize environmental sounds. A commonly adopted pipeline for processing such data involves extracting features to succinctly describe them, modeling the statistics of the features, and deriving predictors that reveal the underlying semantics. The quality of the extracted features is intimately tied to the subsequent stages for obtaining inferences.

A common modus operandi for audio feature extraction is to divide the signal into frames, and extract appropriate features from each frame. Features that reveal the Fourier domain characteristics, and those built on psycho-acoustic principles are typically adopted for representing audio data. However, these features do not often work well for environment sounds. This is evident from the unsatisfactory performance of features such as MFCCs (Mel-frequency cepstral coefficients) and psycho-acoustic features such as pitch, loudness, timbre, etc. in environmental sound recognition [1]. This can be attributed to the differences in the characteristics of natural sounds when compared to conventional specch and audio data. For example, unlike speech signals, which can be modeled as a sequence of phonemes, environmental sounds cannot be atomized into a small set of structures. Similarly, environmental sounds often lack any rhythmic patterns found in music data. On the other hand, methods such as Spectral Dynamic Features (SDF) which attempt to model the temporal behavior of the signals have been shown to provide an improved performance [2]. Alternately, learning features directly from data has been a recent and exciting approach when dealing with large-scale data [3]. For example, deep learning techniques, that can infer a hierarchy of features with increasing complexity, have been successful in modeling natural images [4]. This data-driven approach can be particularly suitable to our case since it can effectively capture the multitudes of variation in the data.

Proposed Approach Overview: In this paper, we develop a new approach for extracting hierarchical features that can be very effective in characterizing natural environmental sounds. Furthermore, we present an approach based on sparse representations to predict tags for novel test samples using the proposed features. In general, data-driven feature learning methods attempt to build meaningful representations by transforming data to a new domain in which the factors relevant to the task in hand are emphasized. This process of adapting features is computationally intensive for large-scale data, and can be severe when learning is carried out in multiple layers. Furthermore, these methods ensure that the representations are invariant to transformations that are not relevant to the desired application. For example, the spatial location of objects in images might not significantly affect the behavior of an object recognition system, and hence it is common to perform spatial aggregation (*pooling*) of local features using histograms, spatial pyramids etc. [5]. Existing strategies for pooling (e.g. max-pooling) do not work for environmental sounds, since capturing the signal dynamics, in both time and frequency, is crucial.

The proposed pipeline for feature extraction and semantic inference is illustrated in Figure I. Instead of using standard audio features, we provide spectrograms as the input to our algorithm. In addition to being effective for modeling environmental sounds, spectrograms enable us to build a more robust pooling strategy. In a deep learning approach, at each layer, filters are inferred from the input, and the

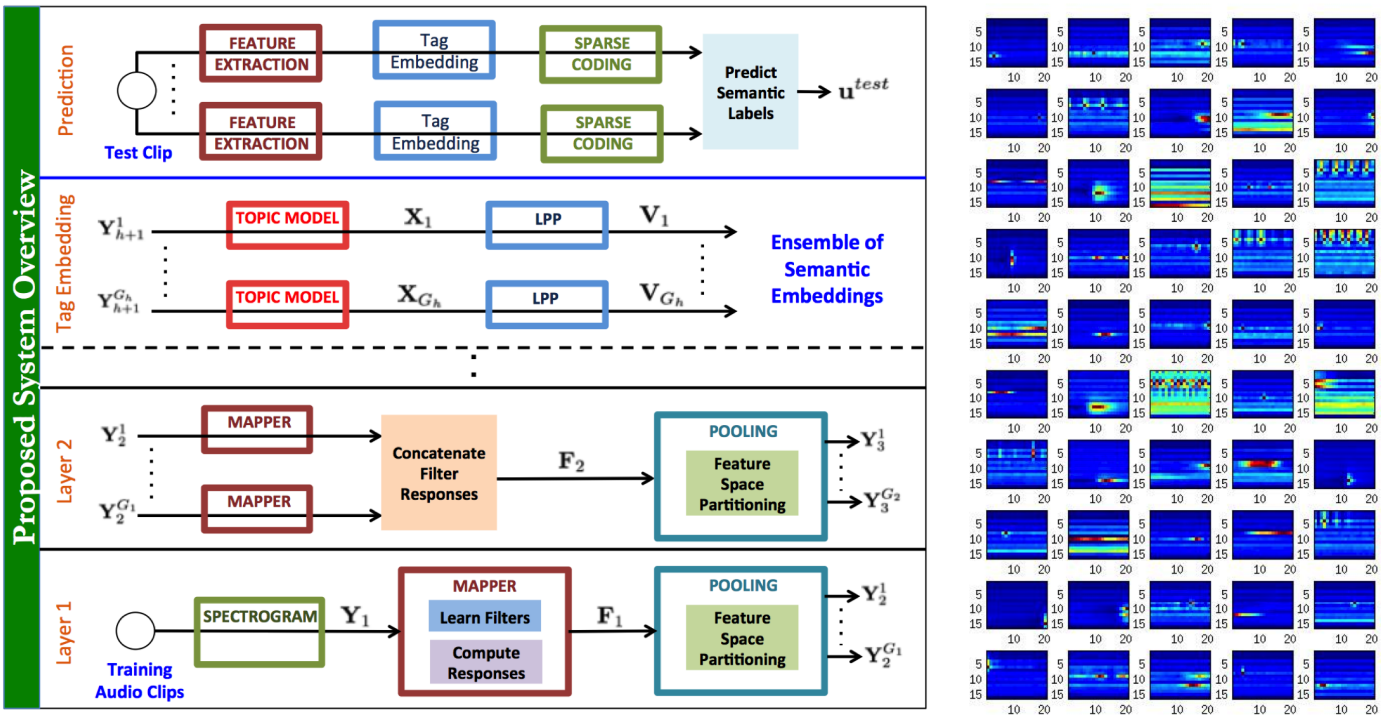

Fig. 1. (*Left*) An overview of the Proposed System. Each layer in the deep feature learning architecture involves a *Mapper* function that infers filters and evaluates responses, and a *Pooling* function that partitions the feature space and aggregates the responses. By constructing an ensemble of semantic embeddings, the underlying tags can be effectively estimated. (*Right*) A subset of layer-1 filters (*dictionary atoms*) learned from the dyadic binned spectrograms of the AASP challenge dataset, where the x- and y-axes correspond to time and dyadic frequency bins respectively.

sparse filter responses are evaluated [6]. We refer to this two step process as the *Mapper* in Figure I. We propose a pooling strategy, *correlation-pooling*, which can preserve the temporal dynamics in the resulting representation. To address the challenges pertinent to scalability, we propose to employ a mini-batch, damped K-hyperline clustering algorithm for inferring filters at each layer. The filters learned are atomic in nature and promote distributed representation of the data-points. Furthermore, we perform partitioning of the feature space based on their temporal similarity so that it is computationally tractable to learn filters in the subsequent layers. Interestingly, the partitioning provides a multi-view representation of the data and we exploit this by building an ensemble of topic models using the pooled features.

Finally, to perform semantic predictions, we construct low-dimensional semantic embeddings for the tag vectors, using graphs based on the topic models. For a test data sample, tag prediction is performed using ℓ_1 reconstruction with the low-dimensional embeddings corresponding to each of the topic models in the ensemble. The final predicted tag vector is obtained as the average of the individual predictions. Experiments with challenging datasets show that the proposed features outperform conventional spectral features used for audio classification. Furthermore, partitioning of the feature space to build an ensemble of representations enables the tag prediction algorithm to capture multiple, possibly unrelated, semantics in the data, simultaneously.

II. Proposed Feature Extraction

In this section, we present a multi-layer feature extraction approach, in which the bottom level consists of building a structure-preserving representation by learning filters from the spectrograms of audio clips (frames). Since preserving temporal dynamics is crucial to characterizing environmental sounds, we will use auto-correlation sequences of the filter responses as the feature for each frame. However, this can be a very high-dimensional vector and hence we perform partitioning of the feature space, based on the temporal structure of the corresponding filter responses. Futhermore, we obtain random projections of the pooled sequences and use the resulting features to train the subsequent layer of the architecture.

Learning Filters: At each layer of the hierarchy, we learn a set of filters, also referred as *dictionary atoms*, using which the data samples are encoded and subsequently aggregated to generate a succinct representation. We develop a modified version of the K-hyperline clustering algorithm (Section 2.2.2, [7]), referred to as mini-batch damped K-hyperline, to infer the dictionary atoms in each level. The input to the first layer of the hierarchy is the dyadic binned spectrograms of frames collected from all audio clips, denoted by the matrix $\mathbf{Y}_1 \in \mathbb{R}^{M_1 \times N_1}$ matrix. The filter responses in a level ℓ are appropriately aggregated to generate the inputs for the next level. The initial dictionary for a level, $\mathbf{D}_\ell^0$, is obtained as K_ℓ randomly chosen input samples (normalized to unit ℓ_2 norm). To combat the challenges with performing clustering on large datasets, we propose to use only a random subset of the input samples for updating the cluster centers during each iteration

TABLE I. MINI-BATCH DAMPED K-HYPERLINE CLUSTERING FOR INFERRING FILTERS IN EACH LAYER OF THE HIERARCHY.

Input:
$\mathbf{Y}_\ell$ - Input matrix of size $M_\ell \times N_\ell$
K_ℓ - Desired number of clusters
T - Number of iterations
B - Size of the mini-batch

Algorithm:
For $t = 1$ **to** T
  Draw a random subset $\mathbf{Y}_\ell^t = [\mathbf{y}_i^t]_{i=1}^B$ from the input matrix $\mathbf{Y}_\ell$.
  Loop for G iterations
    For $i = 1$ **to** B
      - Compute $\mathbf{h}_i = (\mathbf{D}_\ell^{t-1})^T \mathbf{y}_i^t$.
      - Compute $j = \mathbf{argmax}_j(h_{ij})$, where $j = 1 \cdots K_\ell$.
      - Set $\mathbf{u}_j = h_{ij}\mathbf{y}_i + \mathbf{u}_j$.
      - Set $v_j = (h_{ij})^2 + v_j$.
    end
    For $i = 1$ **to** K_ℓ
      - Set $\mathbf{d}_{\ell,j}^t = \frac{\mathbf{u}_j}{v_j} + \mathbf{d}_{\ell,j}^{t-1}$.
      - Normalize $\mathbf{d}_{\ell,j}^t$ to unit ℓ_2 norm.
    end
  end
end

of the algorithm. This procedure is summarized in Table I.

Feature Space Partitioning: Following the dictionary design, we encode all frames using the soft thresholding operator, $\mathbf{f}_{\ell,i} = max\{0, \mathbf{D}_\ell^T \mathbf{y}_{\ell,i} - \alpha\}$, where α is a tuned parameter. Since the filter responses $\{\mathbf{f}_{\ell,i}\}_{i=1}^{N_\ell}$ will be used to construct the input vectors for the subsequent layer, a large K_ℓ can make dictionary learning computationally challenging in the next layer. Consequently, it will be beneficial to partition the feature space into multiple (possibly overlapping) subspaces, and learn an ensemble of filters in the next layer. Though a natural choice is to perform random partitioning of the space, it can be highly suboptimal in describing the temporal dynamics. We propose a greedy partitioning technique that builds multiple $R-$dimensional (possibly overlapping) subspaces, where each subspace is constructed using filters whose responses exhibit similar temporal correlation structures. Let us denote the number of subspaces in layer ℓ as G_ℓ. At a layer ℓ, an audio file indexed by j contains N_ℓ^j frames. Hence, the total number of frames at level ℓ, in the dataset with S clips, can be obtained as $N_\ell = \sum_{j=1}^S N_\ell^j$. The filter responses can now be denoted by the matrix $\mathbf{F}_\ell \in \mathbb{R}^{K_\ell \times N_\ell}$.

In order to partition the feature space, we devise a novel similarity metric for comparing the temporal correlation structures of the filter responses. Let us consider the responses of all N_ℓ^j frames from sound clip j to the filter $\mathbf{d}_{\ell,p}$, denoted by the vector $\mathbf{F}_\ell^{p,j} \in \mathbb{R}^{N_\ell^j}$. We compute $L-$dimensional autocorrelation sequences for $N_{\ell+1}^j$ overlapping windows of $\mathbf{F}_\ell^{p,j}$. By stacking the correlation matrices $\mathbf{C}_\ell^{p,j} \in \mathbb{R}^{L \times N_{\ell+1}^j}$, $\forall j = 1 \cdots S$, we construct the overall correlation matrix $\mathbf{C}_\ell^p \in \mathbb{R}^{L \times N_{\ell+1}}$, where $N_{\ell+1} = \sum_{j=1}^S N_{\ell+1}^j$. The affinity between the correlation structures $\mathbf{C}_\ell^p$ and $\mathbf{C}_\ell^r$ corresponding to the filters $\mathbf{d}_{\ell,p}$ and $\mathbf{d}_{\ell,r}$ is measured as

$$A(p,r) = \frac{\mathbf{Tr}((\mathbf{C}_\ell^p)^T \mathbf{C}_\ell^r)}{\mathbf{Tr}((\mathbf{C}_\ell^p)^T \mathbf{C}_\ell^p) + \mathbf{Tr}((\mathbf{C}_\ell^r)^T \mathbf{C}_\ell^r)}. \quad (1)$$

We adopt a greedy approach to partition the feature space into multiple overlapping subspaces. We begin by randomly choosing a filter and picking $R-1$ filters with the highest similarity measures (from (1)) to form a subspace. Following this, we randomly choose another filter, that has not been used in the earlier step, and construct the second subspace. We repeat this process until all filters are included in atleast one of the subspaces.

Pooling: Since we consider only partially overlapping windows while computing the auto-correlation sequences, the number of frames in a clip j is reduced from N_ℓ^j to $N_{\ell+1}^j$. However, each new frame is now represented by a set of G_ℓ vectors of dimensions $R * L$ each, instead of a single K_ℓ dimensional response vector. Hence, we perform dimensionality reduction on each of the G_ℓ vectors using Random Projections (RP), where the resulting dimension $M_{\ell+1} << R * L$. According to the Johnson-Lindenstrauss lemma [8], RP can approximately preserve isometry in $O(\log(N)/\epsilon^2)$ dimensions for N samples, where ϵ bounds the approximation error. The use of random projections is applicable here as we will further cluster the aggregated representations for learning filters in layer $\ell+1$. The pooled responses, at level ℓ, can now be denoted as $\{\mathbf{Y}_{\ell+1}^g \in \mathbb{R}^{M_{\ell+1} \times N_{\ell+1}}\}_{g=1}^{G_\ell}$, which are the inputs to the next layer.

In the $\ell+1^{th}$ layer, we learn filters, independently, in each of the G_ℓ subspaces identified in the previous layer. The feature space partitioning and the pooling operations significantly reduce the computational cost for the mini-batch clustering. Let $K_{\ell+1}^g$ denote the number of filters learned in the subspace indexed by g ($g = 1 \cdots G_\ell$). The total number of filters at level $\ell+1$ is $K_{\ell+1} = \sum_{g=1}^{G_\ell} K_{\ell+1}^g$ and the filter response matrix, $\mathbf{F}_{\ell+1} \in \mathbb{R}^{K_{\ell+1} \times N_{\ell+1}}$, is obtained by stacking responses from all G_ℓ subspaces. We can then repeat the process of feature partitioning and pooling exactly as the previous layer. In our experiments, we employ a $2-$layer architecture, and the output of our feature extraction process are the matrices denoted by $\{\mathbf{Y}_3^g \in \mathbb{R}^{M_3 \times N_3}\}_{g=1}^{G_2}$. Long duration signals can benefit from more number of layers to further reduce the number of frames in the signal. More depth in the architecture can also be achieved by increasing the overlap length in the windowing operation during pooling. This reduces the rate of decrease in the number of frames for each clip across layers.

III. PREDICTING SEMANTIC LABELS

In this section, we develop an algorithm for predicting semantic tags for environmental sounds based on the features extracted using the proposed approach. Using the pooled features in the multiple partitions, we will learn an ensemble of latent topic models and build topical representations for the audio clips. Following this, we employ a novel strategy based on sparse representations for tag prediction.

Ensemble Topic Models: In order to generate effective representations for audio clips that will enable semantic predictions, we build an ensemble of Replicated Softmax Models (RSM) topic models on the pooled features. Note that, we infer a topic model for each of the subspaces from the feature partitioning. RSM belongs to the family of undirected, energy-based models known as restricted Boltzmann machines (RBM). The visible unit is modeled as a softmax variable instead of a Bernoulli variable as in a conventional RBM. Further details on this technique and its convergence properties

can be found in [9]. Since RSM requires a histogram of words as input, we construct a bag-of-words model based histogram for the audio clips in each of the subspaces. Each of these models can be interpreted of as a local topic model based on a limited set of words, and can reliably predict a subset of tags or categories. Such an ensemble can provide a richer and robust information, and we show that it is particularly suitable for environment sounds which can manifest from numerous categories in real scenarios.

Ensemble Tag Embedding: In order to explore the relationships between the inferred topic models and the ground truth tag vectors (in training data), we compute low-dimensional semantic embeddings for the tag vectors based on graphs constructed using the corresponding topic models. In particular, we assume that the topic features follow a union-of-subspaces model, wherein samples in a subspace can be effectively reconstructed using other samples in that subspace, and construct graphs based on sparse representations. Each of the topic models in the ensemble provides a different semantic embedding that improves the predictability of some tags more than others. Furthermore, we assume that the tag vector for any audio clip is sparse, and that the ℓ_1 reconstruction of a test sample in both the topic feature space and the embedded tag space are similar.

We begin by considering one topic model in the ensemble, and let us denote the set of topic features using the matrix $\mathbf{X} \in \mathbb{R}^{P \times S}$, where P indicates number of topics. The tag (label) vectors for all sound clips are stored in the matrix $\mathbf{U} \in \mathbb{R}^{T \times S}$, where T denotes the total number of labels in the collection used to describe the training set. For the jth clip, the entry $\mathbf{U}_{i,j}$ is set to 1 if the ith tag is associated with that clip. We are interested in computing a linear projection matrix $\mathbf{V} \in \mathbb{R}^{T \times d}$, where $d \ll T$, which will result in similar low-dimensional embeddings for tag vectors with similar topic distributions. The projection directions are estimated using an approach similar to Locality Preserving Projections (LPP) [10], with a sparse coding graph obtained from the topic features. The edge weights in the graph correspond to sparse codes obtained for each feature vector using all other features as the dictionary. This approach is found to be more effective than Euclidean distance based neighborhood in identifying the semantic relationships.

Sparse coding of a topic feature $\mathbf{x}_i$ can be performed as

$$\min_{\boldsymbol{\alpha}_i} \|\mathbf{x}_i - \mathbf{B}\boldsymbol{\alpha}_i\| + \lambda \|\boldsymbol{\alpha}_i\|_1 \tag{2}$$

where the dictionary $\mathbf{B} \in \mathbb{R}^{P \times S-1}$ is designed using all features except $\mathbf{x}_i$ and $\|.\|_1$ denotes the ℓ_1 norm. Denoting the sparse coefficient matrix by $\mathbf{A} \in \mathbb{R}^{S \times S}$, we construct the adjacency matrix for the graph as $\mathbf{W} = |\mathbf{A}| + |\mathbf{A}^T|$.

Given the affinity matrix $\mathbf{W} \in \mathbb{R}^{S \times S}$, we compute the degree matrix $\mathbf{D}$ with each diagonal element containing the sum of the corresponding row or column of $\mathbf{W}$. The d projection directions for LPP are then computed by optimizing

$$\min_{\mathbf{V}} \sum_{i,j=1}^{S} \|\mathbf{V}^T\mathbf{u}_i - \mathbf{V}^T\mathbf{u}_j\|_2^2 w_{ij} \text{ s.t. } \sum_{i=1}^{S} \|\mathbf{V}^T\mathbf{u}_i\|_2^2 d_{ii} = 1.$$

where w_{ij} and d_{ii} are the corresponding elements of the affinity and the degree matrices. This optimization ensures that the embedding preserves the structure defined by the sparse coding based graph. Defining the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$, this optimization problem can be rewritten as

$$\min_{\text{trace}(\mathbf{V}^T\mathbf{U}\mathbf{D}\mathbf{U}^T\mathbf{V})=\mathbf{I}} \text{trace}(\mathbf{V}^T\mathbf{U}\mathbf{L}\mathbf{U}^T\mathbf{V}). \tag{3}$$

and this can be solved using generalized eigen-value decomposition.

Tag Reconstruction: Assuming that our deep architecture for feature extraction contains h levels, for a test audio clip, we extract the features, and construct the G_h (size of the ensemble) topic distribution vectors $\{\mathbf{z}_g\}_{g=1}^{G_h}$. The goal here is to predict the tag vector for the test data using the topic features. For a test clip, we compute the similarity with training clips in terms of features using reconstruction coefficients as

$$\boldsymbol{\beta}_g = \underset{\boldsymbol{\beta}}{\text{argmin}} \|\mathbf{z}_g - \mathbf{X}_g\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{4}$$

Let us denote the low-dimensional embeddings of the training tags corresponding to the gth topic model as $\mathbf{E}_g = \mathbf{V}_g^T\mathbf{U}$. We perform out-of-sample extension for the test sample by estimating its embedding in the semantic space as, $\mathbf{e}_g^{test} = \mathbf{E}_g\boldsymbol{\beta}_g$. Since we know that the tag vector $\mathbf{u}_g^{test}$ is sparse, we solve the following inverse problem to predict the tag vector:

$$\mathbf{u}_g^{test} = \underset{\mathbf{u}}{\text{argmin}} \|\mathbf{e}_g^{test} - \mathbf{V}_g\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1 \tag{5}$$

Finally, we average the predictions from all models in the ensemble to estimate the tag vector: $\mathbf{u}^{test} = \frac{1}{G_h}\sum_{g=1}^{G_h}\mathbf{u}_g^{test}$.

TABLE II. CONFUSION MATRIX FOR THE AASP SCENE CLASSIFICATION DATASET AVERAGED OVER 100 RUNS.

Estimated:	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
A (bus)	**82**	0	11	2	0	0	2	0	2	0
B (busy-street)	0	**100**	0	0	0	0	0	0	0	0
C (office)	5	0	**91**	0	0	1	1	2	0	0
D (open-market)	0	2	0	**87**	0	0	8	2	0	2
E (park)	10	10	36	0	**24**	12	6	0	0	1
F (quiet-street)	2	3	4	8	4	**72**	2	1	2	1
G (restaurant)	6	2	0	17	1	0	**66**	5	0	2
H (supermarket)	18	13	3	5	2	3	15	**20**	1	20
I (tube)	16	12	0	1	0	6	2	1	**56**	8
J (tube-station)	0	12	0	4	1	0	4	9	1	**68**

TABLE III. RECOGNITION PERFORMANCE OF THE PROPOSED APPROACH ON THE PUBLICLY AVAILABLE AASP DEVELOPMENT DATASET IN COMPARISON TO A FEW RELEVANT ENTRIES IN THE AASP SCENE CLASSIFICATION CHALLENGE [11].

Method	**Average Accuracy** (%)
Baseline MFCC + GMM	52
Cochleogram + Tonelikeness	55
Spectral and Temporal and Spatial Features	69
SparseRBM + Uniform Max-pooling	68
SparseRBM + Selective Max-pooling	75
i-vector analysis of MFCC + pLDA	65
Spectral and Temporal features + HMM	72
RQA features + MFCC	71
Our Method	71

IV. EXPERIMENTS

IEEE AASP D-CASE Dataset: This dataset was created for the recently organised IEEE AASP D-CASE scene classification challenge [11]. The dataset consists of 30 seconds long audio clips, recorded using a Soundman binaural microphone (PCM 44.1kHz 16 bit). The publicly available development dataset contains 100 clips with 10 samples from

TABLE IV. ILLUSTRATION OF TAG PREDICTION BEHAVIOR IN THE SUBSPACES FOUND. TABLE SHOWS THE TOP PREDICTED TAGS IN A FEW SUBSPACES FOR A PARTICULAR CLIP. RELATED CONCEPTS ARE CLUSTERED IN A SUBSPACE.

Ground Truth	street-sounds, traffic-noise, sidewalk-conversations, city-streets
SubSpace 1	male, voice, female, vocal, speech
SubSpace 2	traffic, driving, cars, roar, motorbike
SubSpace 3	city, voices, wind, male, station
SubSpace 4	voices, children, screeching, steps, child
SubSpace 5	racing, radio, accelerating, motor, car

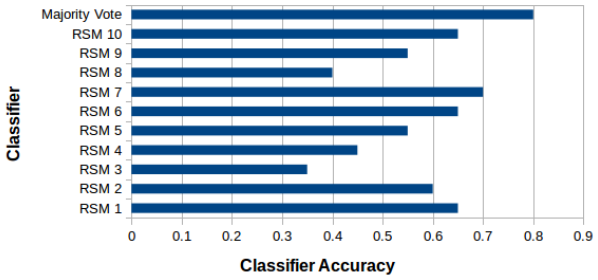

Fig. 2. Effect of using an ensemble of semantic embeddings on the classification performance for a particular run on the AASP challenge dataset. A majority vote classifier outperforms the individual classifiers learned using the corresponding RSM topic models.

each of the following 10 classes: *bus, busy street, office, open market, park, quiet street, restaurant, supermarket, tube, tubestation*. It has an equal balance of indoor and outdoor scenes. We achieve a recognition accuracy of 71%, averaged over 100 runs, when compared to the baseline of 56% obtained using MFCCs. Figure 2 shows the recognition accuracies of each of the RSM topic models, and that of a majority vote classifier combining all of them. Table II shows the confusion matrix summed over 100 runs. The performance can be improved by using a larger dataset to learn filters compared to this dataset. Table III compares our performance to a few related entries from the challenge. The details on the methods used for comparison can be found in [11].

FreeField2010 Dataset: We will demonstrate tag prediction on this, recently released, large dataset containing 7960 audio files (10s long) [12]. These are mostly field recordings, and the total number of tags in the dataset is 7729. The tag *field-recording* was used as the search term to build the database from the freesound website. We use linear SVM for classification, and we report the performance for a subset of tags that have sufficient number of examples in the dataset. The accuracies and f1-scores are illustrated in in Figure 3. The tag prediction behavior of our system in the total space of 7729 tags is illustrated in Table IV. It can be observed that the semantically similar tags have been clustered in subspaces.

V. CONCLUSIONS

To generate features suitable for environment sound recognition we need a pooling method which builds invariance with respect to their dynamic nature. Environment sounds can pack complex semantics in a short time duration. For example, a short clip can comprise of a bird chirping, moving traffic, and people talking. We showed that, it is beneficial to partition the filters into groups, and learn topic models for each group. This ensemble representation provides more flexibility and

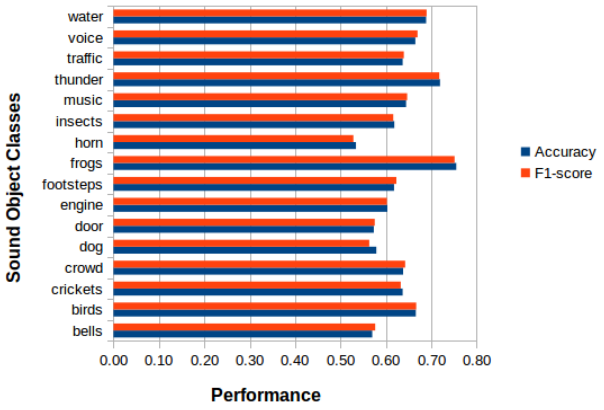

Fig. 3. Prediction performance (accuracy and F1-score) of the proposed approach on a subset of the freefield1010 dataset. This subset was chosen such that it contained a sufficient number of examples for each tag.

robustness to the feature extraction algorithm. By incorporating this novel pooling strategy into a deep architecture, we obtained features that outperform other commonly adopted audio features. Extending this work, it can be beneficial to further understand the complex relationships between the multiple feature subspaces and appropriate semantic concepts. We will also investigate methods to effectively fuse predictions from the ensemble of semantic embeddings.

REFERENCES

[1] S. Chachada and C.-C. J. Kuo, "Environmental sound recognition: A survey," in *IEEE APSIPA*, 2013, pp. 1–9.

[2] M. Karbasi, S. Ahadi, and M. Bahmanian, "Environmental sound classification using spectral dynamic features," in *IEEE ICICS*, 2011, pp. 1–5.

[3] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *IEEE ICASSP*. IEEE, 2013, pp. 8595–8598.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." in *NIPS*, vol. 1, no. 2, 2012.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE CVPR*, vol. 2. IEEE, 2006, pp. 2169–2178.

[6] A. Coates and A. Y. Ng, "Learning feature representations with k-means," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 561–580.

[7] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, "Mixing matrix estimation using discriminative clustering for blind source separation," *Digital Signal Processing*, vol. 23, no. 1, pp. 9–18, 2013.

[8] S. Dasgupta, "Experiments with random projection," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 143–151.

[9] R. Salakhutdinov and G. E. Hinton, "Replicated softmax: an undirected topic model." in *NIPS*, vol. 22, 2009, pp. 1607–1614.

[10] X. He and P. Niyogi, "Locality preserving projections," in *NIPS*, vol. 16, 2003, pp. 234–241.

[11] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *IEEE Workshop on WASPAA*, 2013, pp. 1–4.

[12] D. Stowell and M. D. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," *arXiv preprint arXiv:1309.5275*, 2013.

Prepared by LLNL under Contract DE-AC52-07NA27344.