BLOCK-SPARSE BASIS SETS FOR IMPROVED AUDIO CONTENT ESTIMATION

Sourish Chaudhuri, Rita Singh, Bhiksha Raj

Language Technologies Institute, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA - 15213.
{sourishc, rsingh, bhiksha}@cs.cmu.edu

ABSTRACT

Unsupervised lexicon learning techniques for audio-in-the-wild typically assume that only one of the lexical units is active at any given point in time (hard quantization) or use soft counts to avoid committing to one unit (soft quantization). In reality, the audio will usually be produced as a mixture of the different audio concepts in the lexicon. In this paper, we propose a model where the audio content is assumed to be generated by a mixture of a sparse subset of the lexical units thus guiding the system toward a better estimate of presence of the concepts. We present an approach that builds on current lexicon learning frameworks, and develop a novel algorithm to estimate the contribution of different sources by imposing block-sparsity constraints on the lexicon. Our proposed framework shows significant improvement over the standard lexicon learning framework on a retrieval task for audio-in-the-wild.

***Index Terms*—** block sparsity, concept estimation, audio content analysis, audio retrieval

1. INTRODUCTION

The task of analyzing unconstrained audio (referred to as *in the wild*) for retrieval has recently received a lot of attention. Early approaches to indexing audio data were built around detecting specific sounds in audio streams such as gunshots, laughter, music, crowd sounds etc. [1], or mapping words to acoustic phenomenon [2] using known vocabularies of sounds. In unconstrained audio, the set of such sounds is large, and supervised data is required to build such detectors.

Instead of using audio libraries to build detectors, current approaches attempt to *learn* a lexicon of sounds from the audio data [3, 4] in an unsupervised manner. (Here, and subsequently in this paper, *audio* refers to audio-in-the-wild.) Unlike in speech and music analysis, state of the art systems for large-scale retrieval of audio typically use these lexicons to represent audio data as sequences of the individual lexical units. They assume either that only one lexical unit may be active at any instant (hard quantization) or avoid making a decision by distributing the uncertainty in its estimate among all the different units (soft quantization). Hard quantization will select the dominant source alone, while soft quantization will estimate the likelihood of presence of each of the sources independently instead of jointly.

Typically, multiple different sources may produce sound concurrently, which combine to produce the observed audio. When there are very few atomic sources, learning the lexicon from the data would result in the units modeling mixtures of sources, instead of the atomic sources themselves. However, the number of possible mixture will grow exponentially with the number of true atomic sources, whereas learning and estimation techniques can only use limited vocabularies for computational efficiency. Thus, the very large mixture space will be mapped down to the finite vocabulary of smaller size, collapsing mixtures from different sources together.

In this paper, we extend the current state-of-the-art for analysis of audio in the wild to its logical next step, and present a framework for explicit estimation of mixtures of lexical units in such settings. We model each concept in the lexicon with a set of basis vectors– using such a set allows us to account for various acoustic manifestations of the same concept, by identifying a subspace from which sound is produced for that concept. Each concept source, when active, produces sound using a weighted combination of its basis vectors. The observed audio is assumed to be generated by the additive combination of the sounds produced by the active concepts. We assume, further, that even though there may be many such concepts, only a sparse subset will be active at any given instant. However, since there are no constraints on the number of concept-specific basis vectors that may be active when that concept is active, the weight vector at any instant will be block-sparse.

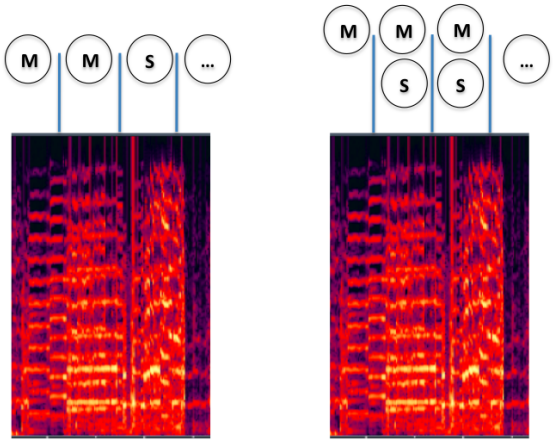

Fig. 1. Audio analysis (L) Only one unit can be active at any time (R) Proposed approach, where a sparse subset of possible concepts can be active concurrently.

To illustrate the difference, consider an example audio from a *birthday party* scene where music is playing and people are talking (Fig. 1). Units corresponding to *music* and *speech* should both be active, but a hard-quantization-based system might choose the dominant music (left panel, Figure 1), while a soft-quantization system would attribute some speech (as well as music) audio to bases for other concepts. This can lead to loss of discriminative information, and makes further analysis using the lexical unit-based representation harder. The proposed framework would be able to recover the

co-occurrence of the different units, resulting in a better interpretation of the audio content, as shown in the right panel of Figure 1.

While we would ideally evaluate our proposed framework on the accuracy with which it estimates source presence, no such datasets currently exist for audio in-the-wild. We evaluated the proposed analysis on an audio retrieval task, where the learnt lexicon and the estimated presence of the units is used to characterize the audio file for retrieval, and obtained significant improvements over standard baselines. However, we expect such models to be useful for various other applications (audio recounting, for instance), since the concurrent estimation of sources allow a finer-level analysis of the audio than current systems would permit.

The rest of the paper is organized as follows: we present prior related work in Section 2, and describe our proposed framework and the various models in Section 3. In Section 4, we present the data and compare performance of the various models on the data, before concluding in Section 5.

2. RELATED WORK

Various approaches have been developed for estimating the content, or the sources responsible for the content of audio. These include the use of sound libraries to initialize a lexicon of sources, as well as the use of unsupervised techniques to learn lexicons from the data itself [1, 2, 3, 4, 5]. Any audio content can then be represented as a sequence of these lexical units, and this discrete representation can be used for indexing, retrieval and classification tasks.

Our framework builds on initial past work in estimating concurrent concept presence in audio content using PLSA [6]. Unlike [6], our approach does not use annotated concept data, since such data is expensive to obtain for audio-in-the-wild, and learns the set of concepts unsupervised. It also incorporates the assumption that only a sparse subset of concepts can be active at any given instant. The ability to estimate these multiple sources allows us to expect more robust performance from this framework. One can think of this proposed framework as imposing a layer of signal separation-based auditory scene analysis [7, 8, 9] with sparsity constraints on top of any of the standard frameworks for audio content estimation. As described in Section 3, the representation of each of the lexical units using a set of bases means that we need to impose sparsity not directly on the weights for the bases but on the sets, using techniques for block-sparse weight estimation [10, 11]. Similar group sparsity-based techniques have been employed for speaker identification [12].

The procedure for estimating the weights of these bases relate to past literature on sparse recovery techniques [13, 14, 15]. It also relates to *dictionary learning* techniques [3, 5, 16], with the difference that we require the learned dictionary to permit block-sparse characterization of data. The process of learning outlined in this paper is similar to techniques used for data decomposition, such as NMF [17] and semi-NMF[18] (where the latter permits the use of negative data and bases), our proposed approach additionally imposes sparsity constraints on the estimation process. Unlike sparse variants of the NMF formulation [19], however, our model requires concept-level sparsity and estimates a block-sparse weight vector, instead.

3. PROPOSED MODEL

The framework proposed in this paper is designed to improve upon one of the limitations of current audio content analysis systems, by allowing multiple sources to be concurrently active. Our generative process assumes, however, that only a sparse subset of all the possible concept sources (dictionary elements) could combine to produce the audio. In this section, we present a novel framework for representing these sources and estimating their presence in the audio.

Our proposed model is designed to improve upon the existing techniques mentioned earlier [3, 5, 4], and begins by using a standard K-means algorithm to learn a dictionary of K units. This dictionary can be used to assign each audio frame to one of the dictionary elements, using Vector Quantization (VQ). Thus, for each dictionary element, a set of audio frames is assigned to it from the VQ estimation step. In our framework, we refer to each of the dictionary elements as an *atomic concept* and model each concept with a set of basis vectors (as opposed to a mean vector for K-means). For each concept, this basis vector set consists of M basis vectors that can be obtained either by randomly sampling exemplar frames (from the set of frames assigned to that concept) or by using an iterative learning process that we will shortly describe.

Let us first introduce the notation used in this section. We assume that the observed data $\mathbf{D}$ (N audio frames, of dimensionality F each; thus, an N x F matrix) has been generated by a non-negative weighted combination of a sparse subset of concepts. We refer to the set of bases for all concepts collectively as $\mathbf{B}$, and that for each concept as $\mathbf{B}_i$ ($i \in [1, K]$, for K concepts). $\mathbf{W}$ refers to the weight matrix of size (KM) x N, with a weight vector for the entire basis set at each time step. The weight for the j-th basis in the i-th concept bag at the t-th time step is indexed by $w_{ij}^{(t)}$.

We first describe the process of estimation of weights given the set of basis vectors for each concept, while imposing concept-level sparsity. Typically, algorithms for sparse estimation apply L_0 norm minimization on the vector being estimated. These include greedy algorithms such as Iterative Hard Thresholding (IHT) [20] and Compressive Sampling Matching Pursuit (CoSaMP) [15]. Alternatively, other approaches relax the NP-hard L_0 minimization problem by using an L_1 penalty instead on the vector, as in the Lasso algorithm [21]. In this paper, we work off of a generalized definition of sparsity for a vector discussed later, which can be shown to be analogous to the L_1 formulation. The generalized definition, however, allows us to measure sparsity on a bounded scale between 0 and 1.

As mentioned before, our approach enforces sparsity at the concept level instead of the individual weights for each basis vector, leading to a block-sparse weight estimation process. To model this, we introduce a coefficient α to measure the activation level of the individual concepts:

$$\alpha_i^{(t)} = \sum_{j=1}^{M} w_{ij}^{(t)}, \forall i \in [1, 2...K] \tag{1}$$

Since the weights are constrained to be non-negative, the activation level is always non-negative. We measure sparsity at the concept level using α as in Equation 2. ϕ represents the concept level of sparsity, and lies between 0 and 1. A higher value for ϕ indicates higher sparsity; ϕ is 1 when only one element in α is non-zero, and is 0 when all elements are equal and non-zero.

$$\phi(\alpha^{(t)}) = \frac{\sqrt{K} - \frac{\sum_i \alpha_i^{(t)}}{\sqrt{\sum_i \alpha_i^{(t)2}}}}{\sqrt{K} - 1} \tag{2}$$

Given a concept dictionary (which includes the set of basis vectors for the concepts), we can estimate a concept-sparse set of weights for the data by optimizing the following objective function

(S represents the desired degree of sparsity):

$$\min_{W} ||D - W^T B||^2 \quad (3)$$
$$s.t. \quad \phi \geq S$$
$$W_i \geq 0, \forall i$$

The objective function above does not have a closed-form solution, but a solution can be obtained using an iterative procedure shown in Algorithm 1. Step 6 in Algorithm 1 requires the projection of the α onto a non-negative space such that the projected vector meets the desired sparsity constraints [19]. The projection operation is described in Algorithm 2.

Algorithm 1 Obtaining an optimal set of weights to satisfy the objective function above, given a set of basis vector bags

Step 1: Initialize **W** randomly
Step 2: Compute α for each observation
Step 3:Project each α vector to be non-negative, have unchanged L2 norm with L1 norm set to achieve desired sparseness
Step 4:$\mathbf{W} \leftarrow \mathbf{W}. * (B^T D)./(B^T B W)$
Step 5:Recompute α based on the new **W**
Step 6:Project each α vector to be non-negative, have unchanged L2 norm with L1 norm set to achieve desired sparseness
Step 7: Go to Step 4, till maximum iterations are reached

Algorithm 2 Projecting a vector (**x**) onto the non-negative space with desired L_1 norm, and unchanged L_2 norm

Step 1: $p_i \leftarrow x_i + (L_1 - \sum_i x_i)/dim(\mathbf{x})$
Step 2: $Z \leftarrow \{\}$
Step 3: If $i \notin Z$, $m_i \leftarrow L_1/(dim(\mathbf{x}) - size(Z))$
Step 4: If $i \in Z$, $m_i \leftarrow 0$
Step 5: $\mathbf{p} \leftarrow \mathbf{m} + \gamma(\mathbf{p} - \mathbf{m})$, where $\gamma \geq 0$ is selected so the resulting **p** satisfies the L_2 norm constraint
Step 6: If $p_i \geq 0, \forall i$, return **p**, end
Step 7: $Z \leftarrow Z \cup \{i : p_i < 0\}$
Step 8: $p_i \leftarrow 0, \forall i \in Z$
Step 9: $c \leftarrow (\sum p_i - L_1)/(dim(\mathbf{x}) - size(Z))$
Step 10: $p_i \leftarrow p_i - c, \forall i \notin Z$
Step 11: Go to Step 3

At training time, the estimated weights can be used to re-estimate the bases. Thus, for the j-th basis for concept i:

$$B_{ij} = \frac{\sum_t w_{ij}^{(t)} \times D^{(t)}}{\sum_t w_{ij}^{(t)}} \quad (4)$$

At test time, the weight vector obtained can be used to estimate the occurrence (F) of the individual concepts in an audio file with T frames:

$$F_i = \sum_{j=1}^{j=M} \sum_{t=1}^{t=T} w_{ij}^{(t)} \quad (5)$$

While the basis vectors for our experiments should ideally be in the spectral domain, the high dimensionality (typically, 257-513) often result in poor basis estimation– indeed, using exemplar-based spectra as an overcomplete basis set is common [22, 23]– and the domain of audio in-the-wild exacerbates this problem. Instead, we work in the dimensionality reduced Mel-Frequency Cepstral Coefficients (MFCC) domain, which has been used in past work to explain MFCCs for speech recognition [23, 24], and we empirically find that this representation proves effective.

4. EXPERIMENTS

Since we do not have labeled data that can be used to directly analyze the accuracy of the estimation process, we evaluate our framework on an audio retrieval task. In this section, we describe the data and the task for our experiments (Section 4.1), the systems that we compare (Section 4.2), the classifier we use (Section 4.3), and finally, the experimental results (Section 4.4).

4.1. Audio Retrieval Dataset and Task

For our experiments, we use the BBC Sound Effects Library CDs 1 – 20 consisting of 1120 different audio clips [25]. This library consists of various conceptual categories of sound, and the audio tracks contain complex audio due to the presence of many different sounds; *e.g.* a supermarket audio contains voices, sound from the checkout bell, trolleys and baskets being stacked. The recordings are of a high and consistent quality, and allow us to compare different systems in a setting where additional confounding factors are not present, as is often the case in Youtube-style, user-generated content with different recording conditions and equipment.

The entire dataset was sampled at 16KHz, and 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC) were extracted and used to represent the data, in all the experiments reported in this paper. While there are a number of categories in this dataset, we only use those that have at least 15 positive instances belonging to the category. Thus, we have the following 10 categories– *Exterior atmospheres*, *Household*, *Interior Backgrounds*, *Transport*, *Animals*, *Audiences*, *Electronic Equipment*, *Water*, *Birds*, *Warfare*. All the other files have a negative label for each of the 10 categories.

The audio retrieval task is defined as follows: given one of the 10 categories as input, the task is to retrieve all audio files belonging to that category from the test collection. We compute *Missed Detection* (MD) and *False Alarm* (FA) rates as follows: suppose there are N_t test files, with C_i belonging to class i, and the detector predicts N_i as belonging to class i, and D_i of these were correct. Then:

$$MD = \frac{C_i - D_i}{C_i}; FA = \frac{N_i - D_i}{N_t - C_i} \quad (6)$$

We report results using the average Area Under MD-v/s-FA Curves (AUC) using 5-fold cross validation on the entire data. Since the curve measures error of the system being evaluated, the lower the area under the curve, the better the performance.

4.2. Systems Used for Retrieval

We use the Vector Quantization technique at the frame level to initialize the set of basis vectors for each concept in our system. The approach presented in Section 3 was then used to compute a concept-sparse estimate of the weights for each audio frame. We then use Equation 5 to compute the relative occurrences of each concept (i):

$$\mathcal{F}_i = \frac{F_i}{\sum_i F_i}, \forall i \in [1, 2...K] \quad (7)$$

The audio file is then represented as a K-dimensional feature vector for the retrieval task, with one feature for each concept where the feature value is the relative occurrence ($\mathcal{F}$) for the concept. We refer to this system using sparse concept representation as **SpaCon**.

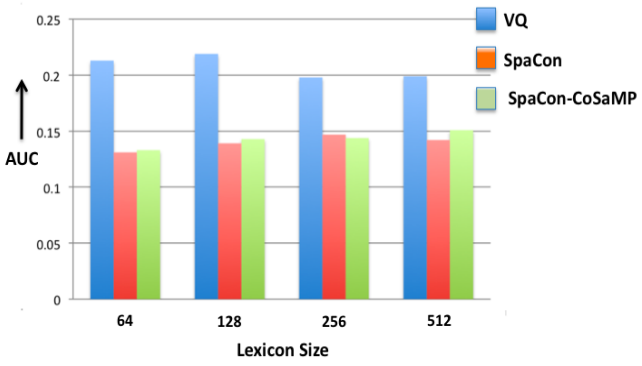

Fig. 2. Comparison of the various systems with average AUC (y-axis) (**lower is better**)

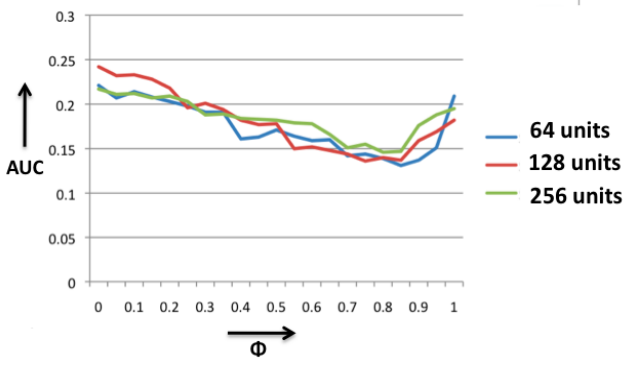

Fig. 3. Effect of changing the desired sparsity on average AUC (y-axis) (**lower is better**) in the SpaCon system

The current state-of-the-art in audio retrieval systems currently use "bag-of-words" representations of recordings generated using the Vector Quantization technique [5, 4]. We use a VQ-based system as the baseline for comparison with our proposed framework. A similar relative occurrence, K-dimensional feature representation is used for the VQ baseline, as well. We refer to this system as **VQ**.

In the estimation process outlined in Algorithm 1, the projection of the concept activation vector (α) to the non-negative space with desired sparsity results in some of the concepts being set to 0 early in the iterative process. The update rule for the weights in Step 4 can no longer recover non-zero weights for those concepts in further iterations. To avoid erroneous concept selection at an early stage, we implement a CoSamp-style approach [15] where the projected vector is augmented with a support set consisting of the $2s$ concepts (for an s-sparse projected vector) with the highest gradient values in each iteration. At the end of the iterations for weight estimation, this augmented vector is finally projected down to the desired sparsity level to obtain the final sparse estimate. Again, audio files are represented using the relative occurrence feature representation, as in the SpaCon system. We refer to this system as **SpaCon-CoSaMP**.

4.3. Random Forest Classifier

The audio retrieval requires us to predict whether each audio file belongs to a particular class or not. Hence, we train binary classifiers for each of the 10 audio categories to predict whether a test file belongs to the class or not (one-versus-all).

The experiments we report in this paper employ a Random Forest [26] classifier for each category. While any classifier could have been used for this task, we chose random forest classifiers as they are resistant to overfitting. Random forests are an extension of decision tree classification techniques, where the training process grows many trees instead of a single one, using held out data is used to get an estimate of the error as trees are added to the forest. The trees in the forest are grown as far as possible, and pruning is not used. Given a new test file, each of the trees in the forest returns a class label, which is used in a weighted vote to determine the final predicted label. In our experiments, we use 500 trees. For details of the training process, the reader is referred to [26].

4.4. Experimental Results

Figure 2 compares areas under the curve (AUC) for the 3 systems described above for varying sizes of the concept lexicon, using 5-fold cross validation on the entire dataset. Recall that since the curve plots missed detections against false alarms, a lower AUC is better.

The sparse concept estimation-based systems provide significant improvement over the VQ baseline. Since the desired degree of sparsity can be modified by the ϕ parameter, this improvement is expected, since the sparsity can be relaxed to create the equivalent of the VQ system. We note that the use of the CoSaMP-style estimation with an augmented support does not appear to improve performance on this retrieval task, resulting in slightly deteriorated performance, in terms of AUC. However, we maintain that such an approach should be optimal in the general case.

The degree of sparsity (ϕ) imposed on the estimation process outlined earlier has a significant effect on the performance of the system. Figure 3 tracks the change in AUC for the SpaCon system with changing the degree of desired sparsity (ϕ) for different lexicon sizes. This plot shows an optimal operating point for ϕ values around 0.85 for the different lexicon sizes.

5. CONCLUSIONS AND FUTURE WORK

We presented a novel, signal-separation based approach to audio content analysis, and demonstrated significant improvement in performance over the commonly used Vector Quantization based technique for audio retrieval on a dataset containing complex audio tracks. While this improvement is exciting by itself, we believe that the importance of this work lies in the fact that it presents a more natural model for the understanding of audio content, due to the assumptions of sparsity among the very large space of natural audio concepts. The improvements in the audio retrieval task suggest that this technique recovers a better estimate of the concept occurrences.

The objective function being optimized in Equation 3 could be modified in appropriate settings to add further constraints. For instance, given prior external knowledge about the relations between the various concepts in the lexicon or about the domain of data, the estimation process could make use of expected structure in estimating the presence of the different concepts.

The improved estimate of concept co-occurrence itself could be used in various ways in the future. Specifically, in past work, we developed a model for extracting patterns over the low-level units (such as VQ) in order to understand how lower-level acoustics (units) combine to produce higher-level semantics [27]. The framework proposed here could be used in conjunction with the one in [27], to better leverage the concurrent occurrence structure for improved semantic analysis. We continue to actively explore these directions.

6. REFERENCES

[1] S.F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. Loui, and J. Luo, "Large-scale multimodal semantic concept detection for consumer video," in *MIR workshop, ACM-Multimedia*, 2007.

[2] M. Slaney, "Mixture of probability experts for audio retrieval and indexing," in *ICME*, 2002.

[3] S. Chaudhuri, M. Harvilla, and B. Raj, "Unsupervised learning of acoustic unit descriptors for audio content representation and classification," in *Interspeech*, 2011, pp. 717–720.

[4] X. Zhuang, S. Tsakalidis, S. Wu, P. Natarajan, R. Prasad, and P. Natarajan, "Compact audio representation for event detection in consumer media," in *Interspeech*, 2011.

[5] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Interspeech*, 2011.

[6] A. Mesaros, T. Heittola, and T. Virtanen, "Latent semantic analysis in sound event detection," in *Proceedings of the European Signal Processing Conference*, 2011.

[7] D. Ellis, "Predication-driven computational auditory scene analysis," *PhD Thesis*, 1996.

[8] A.S. Bregman, "Auditory scene analysis," *International Encyclopedia of the Social and Behavioral Sciences.*, 2004.

[9] Y. Cho and L.K. Saul, "Learning dictionaries of stable autoregressive models for audio scene analysis," in *International Conference on Machine Learning*, 2009.

[10] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, pp. 3042–3054, 2010.

[11] Z. Ben-Haim and Y. C. Eldar, "Near-oracle performance of greedy block-sparse estimation techniques from noisy measurements," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, pp. 1032–1047, 2011.

[12] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Proceedings of Interspeech*, 2012.

[13] S. Becker, J. Bobin, and E.J. Candes, "Nesta: A fast and accurate first-order method for sparse recovery," *SIAM Journal for Imaging Sciences*, vol. 4, pp. 1–39, 2011.

[14] R. Garg and R. Khandekar, "Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property," in *International Conference on Machine Learning*, 2009.

[15] D. Needell and J.A. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, 2009.

[16] M. Yaghoobi, T. Blumensath, and M.E. Davies, "Dictionary learning for sparse approximations with the majorization method," *IEEE Transactions on Signal Processing*, vol. 57, pp. 2178–2191, 2009.

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Neural Information Processing Systems (NIPS)*, 2001.

[18] C.H.Q. Ding, T. Li, and M. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Transactions on Pattern Analysis and Machine Intelligences*, vol. 32, pp. 45–55, 2010.

[19] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[20] T. Blumensath and M.E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 629654, 2008.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, pp. 267–288, 1994.

[22] B.Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proceedings of Interspeech*, 2010.

[23] T. N. Sainath, D. Nahamoo, D. Kanevsky, B. Ramabhadran, and P. M. Shah, "Exemplar-based sparse representation phone identification features," in *Proceedings of ICASSP*, 2011.

[24] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representation features for speech recognition," in *Proceedings of Interspeech*, 2010.

[25] "Bbc sound effects library original series, http://www.sound-ideas.com/bbc.html," .

[26] L. Breiman, "Random forests," *Machine Learning*, vol. 45, 2001.

[27] S. Chaudhuri and B. Raj, "Unsupervised structure discovery for semantic analysis of audio," in *Neural Information Processing Systems*, 2012.