MODIFIED LASSO SCREENING FOR AUDIO WORD-BASED MUSIC CLASSIFICATION USING LARGE-SCALE DICTIONARY

Ping-Keng Jao, Chin-Chia Michael Yeh and Yi-Hsuan Yang

Research Center for Information Technology Innovation, Academia Sinica, Taiwan

ABSTRACT

Representing music information using audio codewords has led to state-of-the-art performance on various music classification benchmarks. Comparing to conventional audio descriptors, audio words offer greater flexibility in capturing the nuance of music signals, in that theoretically each codeword can be considered as a quantization of the music universe, and that the quantization goes finer as the size of the dictionary (i.e., audio codebook) increases. In practice, however, the size of the dictionary is usually limited, as codeword assignment is a computationally intensive process. To facilitate the use of a large vocabulary dictionary, this paper presents two modifications of a LASSO screening technique developed in the compressive sensing field to reduce the time cost in codeword assignment. The first modification exploits the repetition nature of music signals, whereas the second one relaxes a screening constraint that is specific to reconstruction but not for classification. Our experiments show that, with the proposed algorithm, we can afford using a dictionary of 10,000 codewords with runtime close to the case of using a dictionary of 1,000 codewords. Moreover, using the larger dictionary significantly improves the mean average precision (MAP) of music genre tagging from 0.176 to 0.211, for tagging thousands of tracks with 147 possible genre tags.

Index Terms— Sparse coding, feature learning, LASSO screening, music information retrieval, genre classification

1. INTRODUCTION

Feature learning is a burgeoning research topic in the broad signal processing and machine learning communities [1]. For music information retrieval (MIR), representing music information as a term-document structure comprising of elementary audio codewords have been found competitive or even superior to conventional, hand-crafted audio features [2–5].

Given a dictionary (audio codebook) \( D \in \mathbb{R}^{m \times k} \), which is a finite collection of \( k \) codewords \( d_j \in \mathbb{R}^m \), \( j = 1, \ldots, k \), an input acoustic feature vector \( x \in \mathbb{R}^m \) can be replaced by a linear combination of the codewords, leading to the so-called \textit{audio word} (AW), or \textit{bag-of-frames} representation [5–10]. Specifically, the AW representation is a \( k \)-dimensional vector \( \alpha \in \mathbb{R}^k \) computed by a \textit{codeword assignment} function \( \alpha = f(D, x) \). For the case of \( k \gg m \), it has been shown that the following \textit{sparse coding} (SC) formulation empirically performs better than competing methods such as the classic vector quantization (VQ) method [6, 11],

\[
\alpha^* = f_{SC}(D, x, \lambda_c) = \arg \min_{\alpha} \frac{1}{2} \|x - D\alpha\|^2_2 + \lambda_c \|\alpha\|_1. \tag{1}
\]

This formulation leads to codes that are sparse in support (i.e., small \( \|\alpha\|_1 \)) but are sufficient to reconstruct or to interpret the input signal (i.e., small \( \|x - D\alpha\|^2_2 \)); the parameter \( \lambda_c \) controls the balance between these two terms. This problem can be efficiently solved by for example the least angle regression (LARS)-LASSO algorithm [12] given a dictionary of moderate size. Because the AW representation is high-dimensional but sparse, using linear support vector machine (SVM) [13] for training audio classifiers has been found effective [9, 11].

The dictionary \( D \) is usually generated from a possibly unlabeled music collection \( \mathcal{X} \) in off-line [4, 5]. As each codeword \( d_j \) represents a sample of the acoustic space \( \mathcal{U} \subseteq \mathbb{R}^m \), theoretically \( D \) can represent every nuance of music signals if 1) \( \mathcal{X} \) is representative of the music universe \( \mathcal{U} \) and 2) sufficiently large number of codewords are drawn from \( \mathcal{X} \). The first condition can be approached by generating the dictionary from for example the million song dataset [2, 14, 15], a

\begin{align*}
\text{(a)} & \quad \lambda = 3.5\lambda_0 \\
\text{(b)} & \quad k = 10000
\end{align*}

Fig. 1. Time cost of performing LARS-LASSO sparse coding [12] with (red line with circles) or without (blue dashed line) the proposed modified LASSO screening (a) for fixed \( \lambda_s = 3.5\lambda_0 \) and varying dictionary size \( k \) (from 250 to 10,000) and (b) fixed \( k=10,000 \) and varying \( \lambda_s \) (cf. Section 4.1).
the publicly-available collection of audio features for a million contemporary popular music tracks. To meet the second condition, a straightforward approach is to increase the vocabulary size by drawing more samples from \( X \). In practice, however, the approach is not feasible as SC using large vocabulary can be exceedingly time consuming [16].

The dotted line in Fig. 1(a) shows the time cost of encoding the 12-bin chromagram of a song (i.e., \( m = 12 \)) using LARS-LASSO on a regular PC. It can be found that the runtime grows linearly along with \( k \), and that it costs almost 5 seconds for the case of \( k = 10,000 \). Better efficiency is desirable for large-scale or mobile applications.

In this paper, we show that we can reduce the runtime remarkably (as shown by the solid line in Fig. 1(a)) with a modified LASSO screening technique without compromising the effectiveness of the AW representation. This amounts to ten-fold speed increase for \( k = 10,000 \). LASSO screening is a novel technique proposed in the compressive sensing field to filter redundant codewords for a given input without affecting the existence of optimal solution [16–18]. We show that existing screening technique is less applicable to music signal processing and propose two modifications. Our evaluation shows that the proposed modifications effectively speed up SC and improves the accuracy of an AW-based music genre classification system.

Accelerating AW-based music classification has been studied recently. For example, Yeh et al. [19] employed dimension reduction techniques to reduce \( m \) and multi-frame or temporal sampling techniques to reduce the number of frames to be encoded for an audio signal. Yang [20] investigated the use of randomized clustering forest (RCF) to replace SC. On the other hand, for deep learning techniques distributed architecture is usually employed [2, 4, 21]. The present work differs from the prior arts in the following aspects. First, we focus on a system that can be built on a single machine. Second, the proposed methods are designed specifically for SC, which usually leads to more effective AW representation than competing methods such as RCF or VQ [11, 20]. Lastly, this work represents an early attempt that applies screening to improve the efficiency of SC with a large-scale dictionary. In contrast, the dictionary size \( k \) used in previous work is usually less than 2,000 [4, 5, 10, 15].

2. MODIFIED LASSO SCREENING

2.1. LASSO screening

The principle idea of LASSO screening is to remove code-words in \( D \) that are unlikely to be useful for encoding a specific signal \( x \), before actually performing SC [16–18]. From the dual form of Eq. 1,

\[
\theta^* = \arg \max_{\theta} \frac{1}{2} \| x \|^2 - \frac{\lambda^2}{2} \| \theta - \frac{x}{\lambda_s} \|^2, \\
\text{s.t.} \quad |\theta^T d_j| \leq 1 \quad \forall j = 1, \ldots, k,
\]

**Algorithm 1** Modified LASSO screening based SC

**Input:** \( \lambda_c; \lambda_s; X \in \mathbb{R}^{m \times n} := [x_1, \ldots, x_n] \);

**Output:** \( A \in \mathbb{R}^{k \times n} := [a_1, \ldots, a_n] \);

1: \( \Omega \leftarrow \emptyset \);
2: for \( t = 1 \) to \( n \) do
3: \( w_t \leftarrow \text{screening}(D, x_t, \lambda_s) \);
4: \( \Omega \leftarrow \Omega \cup w_t \);
5: end for
6: for \( t = 1 \) to \( n \) do
7: \( a_t \leftarrow f_{\text{SC}}(\Omega, x_t, \lambda_c) \);
8: end for

rules for irrelevant codeword removal can be discovered by exploring the relation between the dual variable \( \theta \) and the primal variable \( \alpha \). Given that the input \( x \) and each \( d_j \) are normalized to unit energy (i.e., \( x^T x = 1 \)), Xiang et al. deduced a method called the “sphere test–3” to select codewords from \( D \) for encoding \( x \) [16, 17]. Specifically, the test calculates the \( x^T d_j \), the correlation between the input and a codeword, and removes \( d_j \) if the following inequality holds,

\[
|x^T d_j - (\lambda_{\text{max}} - \lambda_s) d_j^T d_j| < \lambda_s \left( 1 - \sqrt{\lambda_{\text{max}} - 1} \left( \frac{\lambda_{\text{max}}}{\lambda_s} - 1 \right) \right),
\]

where \( \lambda_{\text{max}} \equiv \max_j |x^T d_j| \) is the largest absolute correlation between the input and the codewords, \( d_\ast \in \{ \pm d_j \}_{j=1}^k \) is the codeword that leads to \( \lambda_{\text{max}} \), and \( \lambda_s \leq \lambda_{\text{max}} \) is a parameter. After applying the above test to all the codewords, we obtain a dictionary subset \( w \in \mathbb{R}^{m \times k'} \), \( k' < k \), for \( x \).

\[
w = \text{screening}(D, x, \lambda_s).
\]

It has been proved that, when \( \lambda_s = \lambda_c \) the screening function will not affect the optimal solution of SC [16].

In practice, different \( w \) are used for different inputs. To make the dimension of the coding result consistent, we fill zeros to the entries corresponding to removed codewords to increase the dimension from \( k' \) to \( k \).

2.2. Modifications

For a song with \( n \) frames, we have to perform SC for each frame \( x_t \) individually. In consequence, we would also have to use LASSO screening \( n \) times to compute \( w_t \) for each \( x_t \). However, it might not be necessary to encode each \( x_t \) with possibly a nearly totally different dictionary, given the repetition nature of music (i.e., the frames might be similar with one another). In addition, using different subsets of \( D \) to encode a song may suffer from either extra memory transfer or memory discontinuity problems. Our pilot study shows that the extra cost in memory transfer can be an important issue — we compared runtime of SC without screening and SC with any of the existing screening techniques [16, 17] and found no acceleration but severe speed increase in many cases.
3. EXPERIMENT SETUP & SYSTEM OVERVIEW

Two datasets were utilized in the performance study: the million song dataset (MSD) [14] for dictionary generation and the CAL10k dataset [23] for evaluating the accuracy of AW-based music classification. The CAL10k dataset contains the annotation over 147 genre tags for 10,870 songs made by expert musicologists from Pandora (http://www.pandora.com).

For the low-level feature representation \( x \), we used the 12-D timbre descriptors (ENT) and 12-D pitch descriptors (ENP) computed by the EchoNest API (http://developer.echonest.com/). ENT describes the timbre characteristics of the magnitude spectrogram, whereas ENP is chroma-like [23]. The features of a song are not in the frame-level but in the segment-level computed by the EchoNest API, where each segment corresponds to an acoustically homogenous fragment with multiple frames. In consequence, LARS-LASSO was employed to encode the segment-level features.

We respectively generated two dictionaries for ENT and ENP by using the “exemplar-based” approach, randomly selecting one segment-level feature vector from \( k \) random songs of MSD and considering the union of the vectors as a dictionary of size \( k \). The same \( k \) is used for timbre and pitch. This approach bypasses the need of using dictionary learning algorithms [1, 22, 24], which is not the focus of this paper. Moreover, exemplar-based dictionary has been found effective for music genre classification [20, 25].

For classification, we computed \( \alpha^{\text{ENT}}_t \in \mathbb{R}^k \) and \( \alpha^{\text{ENP}}_t \in \mathbb{R}^k \) for each segment and used the concatenation \( \hat{\alpha}_t = [\alpha^{\text{ENT}}_t; \alpha^{\text{ENP}}_t] \in \mathbb{R}^{2k} \) as the final segment-level feature. To obtain the song-level representation, which is used as input to SVM, we performed sum-pooling over a song by \( \alpha^{\text{pool}} = \sum_{t=1}^n \hat{\alpha}_t \). In addition, \( L_1 \) and cubic-root power normalization were applied to \( \alpha^{\text{pool}} \) for better performance [5].

Fig. 2 depicts the diagram of the proposed system. In our implementation, we cast the multilabel tagging problem into binary classification problems and used LIBLINEAR [13] for classifier training and prediction. The SVM parameter \( C \) was fixed to 8 throughout the experiments.

4. RESULT

4.1. Efficiency of SC with modified LASSO screening

We first present a preliminary efficiency evaluation of SC with or without screening, using a fixed \( \lambda_s = \lambda_c = 1/\sqrt{12} \) [10] and varying \( \lambda_s \) and \( k \), for encoding a song randomly selected from CAL10k. On average, each song in CAL10k has 840 segments, so we have to perform nearly a thousand SC (and
screening) for a song. Fig. 1(a) clearly shows that the proposed LASSO screening (with $\lambda_s=3.5\lambda_0$) greatly reduces the runtime for SC; the speed up gets more pronounced as $k$ increases. This result shows that screening is important for a large-scale dictionary. On the other hand, Fig. 1(b) shows that LASSO screening is effectively only when sufficiently large $\lambda_s$ is used; if the original screening technique is applied (i.e., setting $\lambda_s = \lambda_0$), there is no speed up.

Table 1 shows how song-level fusion affects the efficiency. The following three important observations are made. First, the conventional LASSO screening (i.e., $\lambda_s = \lambda_0$ and without union) does not speed up but slow down the runtime to 0.03x, comparing to the runtime of SC without any screening. Second, higher $\lambda_s$ leads to more speed up, as expected. Third, song-level union boosts the efficiency further. For $\lambda_s = 3.5\lambda_0$ and $k = 10,000$, the speed up attains 8.03x.

However, better efficiency in SC might come with lower accuracy of the resulting AW representation. To study this, we report the accuracy for genre tagging based on the proposed SC features, using the extreme case $\lambda_s = 3.5\lambda_0$.\(^1\)

### 4.2. Accuracy of AW-based genre tagging

The accuracy of genre tagging for CAL10k was computed by averaging the result of the five train/test splits specified by Tingle et al. [23]. The following four measures are computed: the area under the receiver operating characteristic curve (AUC), mean average precision (MAP), 10-precision (10-Prec) and R-precision (R-Prec), according to [23].

The upper half of Table 2 shows the performance of SC features without (w/o) and with the modified LASSO screening as a function of dictionary size $k$. It can be found that better accuracy is obtained as $k$ increases, as expected. The performance improvement is pronounced for MAP, 10-Prec and R-Prec. This result supports our interest in using a large-scale dictionary. From Table 2, we also see that screening does not affect the accuracy much. Although slightly better accuracy can be obtained without screening, the modified LASSO screening does not trade much accuracy for efficiency.

The lower half of Table 2 shows the result of two baseline methods (a random baseline and linear SVM using the concatenated ENT and ENP features without SC) and three existing methods [19,23] — the first two existing methods [19] applied SC on the raw magnitude spectrum (instead of EchoNest features) and employed dimension reduction and multi-frame representation, whereas the last one [23] used Gaussian Mixture Model (GMM) to model the 36-D ‘ENT+Δ’ feature with first- and second-order temporal derivatives. The dictionary size $k$ for SC was set to 1,024 in [19] due to the computational cost of SC.\(^2\) We can find that the proposed SC features are competitive comparing with these prior arts. Remarkably higher accuracy in 10-Prec and R-Prec is obtained.

### 5. CONCLUSIONS

In this paper, we have presented an efficient and effective sparse coding engine for MIR. It adds two modification to a LASSO screening method to enable the use of a large dictionary, which in turn improves the ability of the codewords in representing the fine details of music signals. The first modification (song-level union) exploits the repetitive nature of music to reduce memory overhead, and the second one relaxes a constraint that is not required for classification. A large-scale dictionary generated from the million song dataset leads to state-of-the-art result in a genre tagging benchmark, the CAL10k dataset. In addition, our study provides empirical evidence showing larger dictionary is indeed favorable.

---

\(^1\)In effect this sets $\lambda_s = 3.5/\sqrt{12} > 3.5 > 1$, the maximal possible value for $\lambda_{\text{max}}$. A prerequisite of Eq. 3 is $\lambda_s \leq \lambda_{\text{max}}$ [16].

\(^2\)Moreover, actually a slightly different subset of CAL10k was employed by Yeh et al. [19], rendering the result not exactly comparable.
6. REFERENCES


