Finding Musically Meaningful Words Using Sparse CCA

David A. Torres, Douglas Turnbull, Bharath K. Sriperumbudur,
Luke Barrington & Gert Lanckriet

University of California, San Diego

Introduction

Goal: Create a content-based music search engine for natural language queries.

- it annotates songs with semantically meaningful words and retrieve relevant songs based on a text query.
- CAL music search engine [Turnbull et al., 2007].

Introduction

Goal: Create a content-based music search engine for natural language queries.

- it annotates songs with semantically meaningful words and retrieve relevant songs based on a text query.
- CAL music search engine [Turnbull et al., 2007].

Problem: Picking a vocabulary of musically meaningful words (vocabulary selection).

- discover words that can be modeled accurately.

Introduction

Goal: Create a content-based music search engine for natural language queries.

- it annotates songs with semantically meaningful words and retrieve relevant songs based on a text query.
- CAL music search engine [Turnbull et al., 2007].

Problem: Picking a vocabulary of musically meaningful words (vocabulary selection).

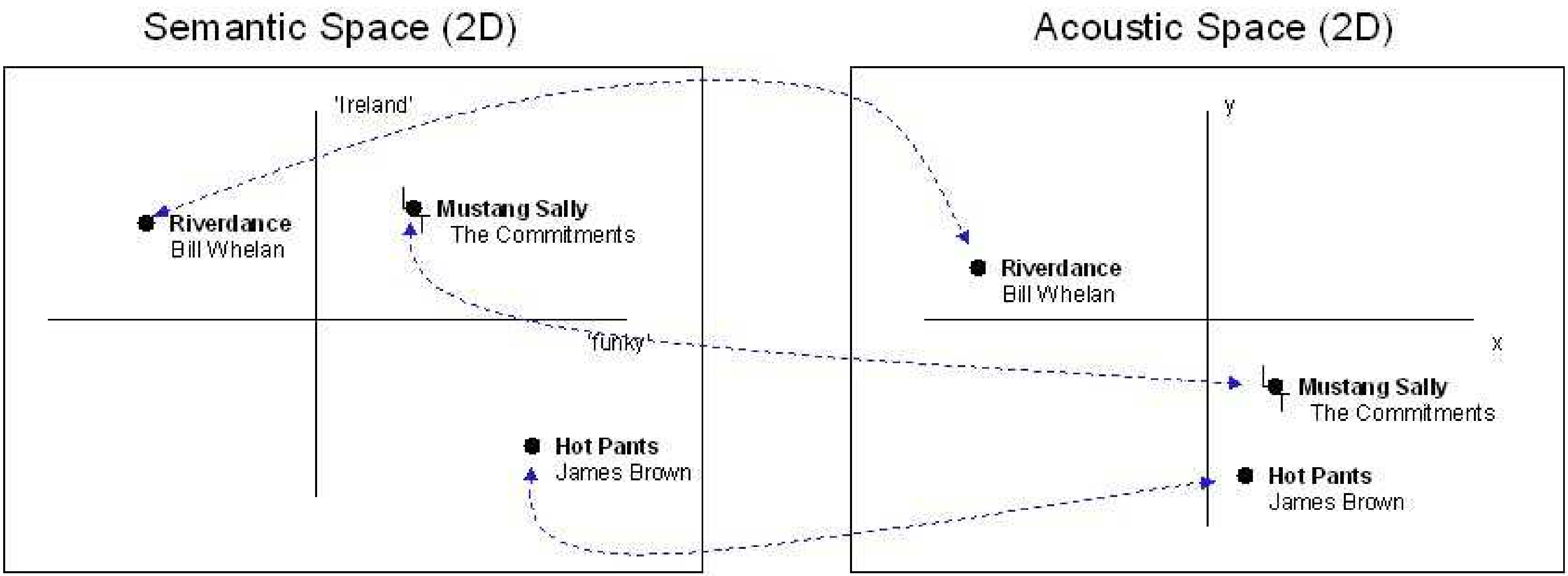
- discover words that can be modeled accurately.

Solution: Find words that have a high correlation with the audio feature representation.

Two-view Representation

Consider a set of annotated songs. Each song is represented by:

- Annotation vector in a semantic space
- Audio feature vector in a acoustic space



Semantic Representation

Vocabulary of words

- CAL 500: 174 phrases from a human survey
  - Instrumentation, genre, emotion, usages, visual characteristics
- LastFM: 15,000 tags from social music site
- Web mining: 100,000+ words minded from text documents

Annotation vector, $\mathbf{s}$

- Each element represents the semantic association between a word and the song.
- $\mathbf{s} \in \mathbb{R}^d$, where d is the size of the vocabulary.
- Example: Frank Sinatra's "Fly me the moon"
  - Vocabulary={funk, jazz, guitar, female vocals, sad, passionate}
  - $\mathbf{s} = [\frac{0}{4}, \frac{3}{4}, \frac{4}{4}, \frac{0}{4}, \frac{2}{4}, \frac{1}{4}]$

Acoustic Representation

Each song is represented by an audio feature vector **a** that is automatically extracted from the audio-content.

- Mel-frequency cepstral coefficients (MFCC).

Canonical Correlation Analysis

Let $X \in \mathbb{R}^{d_x}$ and $Y \in \mathbb{R}^{d_y}$ be two random variables.

Problem: Find $\mathbf{w}_x$ and $\mathbf{w}_y$ such that $\rho(\mathbf{w}_x^T\mathbf{X}, \mathbf{w}_y^T\mathbf{Y})$ is maximized.

Solution: Solve

$$\max_{\mathbf{w}_x,\,\mathbf{w}_y} \frac{\mathbf{w}_x^T\mathbf{S}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{S}_{xx}\mathbf{w}_x}\sqrt{\mathbf{w}_y^T\mathbf{S}_{yy}\mathbf{w}_y}} \tag{1}$$

which is equivalent to

$$\max_{\mathbf{w}_x,\,\mathbf{w}_y} \quad \mathbf{w}_x^T\mathbf{S}_{xy}\mathbf{w}_y$$
$$\text{s.t.} \quad \mathbf{w}_x^T\mathbf{S}_{xx}\mathbf{w}_x = 1\,,\,\mathbf{w}_y^T\mathbf{S}_{yy}\mathbf{w}_y = 1. \tag{2}$$

The above is the variational formulation of CCA.

Canonical Correlation Analysis

- In our analysis, a variation of Eq. (2) is used as given below.

$$\max_{\mathbf{w}} \quad \mathbf{w}^T\mathbf{P}\mathbf{w}$$
$$\text{s.t.} \quad \mathbf{w}^T\mathbf{Q}\mathbf{w} = 1. \tag{3}$$

where $\mathbf{P} = \begin{pmatrix} \mathbf{0} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{0} \end{pmatrix}$, $\mathbf{Q} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{yy} \end{pmatrix}$ and $\mathbf{w} = \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix}$.

- Eq. (3) is a generalized eigenvalue problem with $\mathbf{P}$ being indefinite and $\mathbf{Q} \in \mathbb{S}_{++}^{d_x+d_y}$.

Need for sparsity

- CCA solution is usually not sparse.
  - The solution vector has components along all the features (here, words).
  - Difficult to interpret the results.
- Few relevant features might be sufficient to describe the correlation.
- In our application, vocabulary pruning results in modeling fewer words.

Solution: Sparsify the CCA solution.

Sparse CCA

Heurisitc: $\mathbf{w}_y = [w_{y_1}, \ldots, w_{y_{n_y}}]^T$. If $|w_{y_i}| < \epsilon$, choose $w_{y_i} = 0$. (non-optimal)

Solution: Introduce the sparsity constraint in CCA's variational formulation.

Sparse CCA: The variational formulation is given by

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}^T \mathbf{P} \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{Q} \mathbf{w} = 1 \\ & \|\mathbf{w}\|_0 \le k, \end{aligned} \tag{4}$$

where $1 \le k \le n$, $n = d_x + d_y$ and $\|\mathbf{w}\|_0$ is the cardinality of $\mathbf{w}$.

Issues: Eq. (4) is NP-hard and therefore intractable. ℓ_1-relaxation is still computationally hard.

Convex Relaxation

Primal:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}^T \mathbf{P} \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{Q} \mathbf{w} \leq 1 \\ & \|\mathbf{w}\|_1 \leq k. \end{aligned} \tag{5}$$

Trick: Compute the bi-dual (dual of the dual of the primal).

Bi-dual:

$$\begin{aligned} \max_{\mathbf{W},\mathbf{w}} \quad & \mathrm{tr}(\mathbf{W}\mathbf{P}) \\ \text{s.t.} \quad & \mathrm{tr}(\mathbf{W}\mathbf{Q}) \leq 1 \\ & \|\mathbf{w}\|_1 \leq k \\ & \begin{pmatrix} \mathbf{W} & \mathbf{w} \\ \mathbf{w}^T & 1 \end{pmatrix} \succeq 0. \ \ (SDP) \end{aligned} \tag{6}$$

Issue: SDP relaxation is prohibitively expensive to solve for large n.

Approximation to $||\mathbf{x}||_0$

- Two observations
  - The ℓ_1-norm relaxation does not simplify Eq. (4) $\Rightarrow$ a better approximation to cardinality would improve sparsity.
  - The convex SDP approximation to Eq. (4) scales terribly in size $\Rightarrow$ use a locally convergent algorithm with better scalability.
- Eq. (4) can be written as

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}^T\mathbf{P}\mathbf{w} - \rho\,||\mathbf{w}||_0 \\ \text{s.t.} \quad & \mathbf{w}^T\mathbf{Q}\mathbf{w} \leq 1, \end{aligned} \tag{7}$$

  where $\rho \geq 0$.
- Approximate $||\mathbf{x}||_0$ by $\sum_{i=1}^{n} \log(|x_i|)$. (Refer to [Sriperumbudur et al., 2007] for more details)

Approximation to $||\mathbf{x}||_0$

- Eq. (7) can be written as

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mu||\mathbf{w}||^2 - \left(\mathbf{w}^T(\mathbf{P}+\mu\mathbf{I})\mathbf{w} - \rho\sum_{i=1}^{n}\log|w_i|\right) \\ \text{s.t.} \quad & \mathbf{w}^T\mathbf{Q}\mathbf{w} \le 1. \end{aligned} \tag{8}$$

  where $\mu \ge \max(0, -\lambda_{min}(\mathbf{P}))$.

- The objective in Eq. (8) is a difference of two convex functions and therefore is a d.c. program.

- Solving Eq. (8) using the DC minimization algorithm (DCA) [Tao and An, 1998] yields the following algorithm.

Sparse CCA Algorithm

Require: $\mathbf{P} \in \mathbb{S}^n$, $\mathbf{Q} \in \mathbb{S}^n_{++}$ and $\rho \geq 0$

1: Choose $\mathbf{w}_0 \in \{\mathbf{w} : \mathbf{w}^T\mathbf{Q}\mathbf{w} \leq 1\}$ arbitrarily

2: **repeat**

3:

$$\begin{aligned} \bar{\mathbf{w}}^* = \arg\min_{\bar{\mathbf{w}}} \quad & \mu\bar{\mathbf{w}}^T\mathbf{D}^2(\mathbf{w}_l)\bar{\mathbf{w}} - 2\mathbf{w}_l^T[\mathbf{P} + \mu\mathbf{I}]\mathbf{D}(\mathbf{w}_l)\bar{\mathbf{w}} + \rho||\bar{\mathbf{w}}||_1 \\ \text{s.t.} \quad & \bar{\mathbf{w}}^T\mathbf{D}(\mathbf{w}_l)\mathbf{Q}\mathbf{D}(\mathbf{w}_l)\bar{\mathbf{w}} \leq 1 \end{aligned} \tag{9}$$

4: $\quad \mathbf{w}_{l+1} = \mathbf{D}(\mathbf{w}_l)\bar{\mathbf{w}}^*$

5: **until** $\mathbf{w}_{l+1} = \mathbf{w}_l$

6: **return** $\mathbf{w}_l$, $\bar{\mathbf{w}}^*$

where $\mathbf{D}(\mathbf{w}) = \text{diag}(\mathbf{w})$.

- solves a sequence of convex quadratically constrained quadratic programs (QCQPs).

Modification to Vocabulary Selection

- For vocabulary selection, the sparsity constraint is required only on $\mathbf{w}_y$ instead of on $\mathbf{w}$.
- Modify Eq. (9) as

$$\begin{aligned} \bar{\mathbf{w}}^* = \arg\min_{\bar{\mathbf{w}}} \quad & \mu \bar{\mathbf{w}}^T \mathbf{D}^2(\mathbf{w}_l)\bar{\mathbf{w}} - 2\mathbf{w}_l^T[\mathbf{P} + \mu\mathbf{I}]\mathbf{D}(\mathbf{w}_l)\bar{\mathbf{w}} + \|\boldsymbol{\tau} \circ \bar{\mathbf{w}}\|_1 \\ \text{s.t.} \quad & \bar{\mathbf{w}}^T \mathbf{D}(\mathbf{w}_l)\mathbf{Q}\mathbf{D}(\mathbf{w}_l)\bar{\mathbf{w}} \le 1 \end{aligned} \tag{10}$$

where $(\mathbf{p} \circ \mathbf{q})_i = p_i q_i$ and $\boldsymbol{\tau} = [0, 0, \overset{d_x}{\ldots}, 0, \rho, \rho, \overset{d_y}{\ldots}, \rho]^T$.

- The non-zero elements of $\mathbf{w}_y$ can be interpreted as those words which have a high correlation with the audio representation.

Setting ρ: Not straightforward (increasing ρ reduces the vocabulary size).

Issues: Quality of the solution is hard to derive unlike in SDP.

Experimental Setup

Dataset: CAL500 [Turnbull et al., 2007]

- 500 songs by 500 artists

Semantic representation:

- 173 words (e.g. genre, instrumentation, usages, emotions, vocals, etc.)
- Annotation vector, $\mathbf{s}$ is an average from 4 listeners.
- Word agreement score: measures how consistently listeners apply a word to songs.

Acoustic representation:

- Bag of dynamic MFCC vectors (52-dimensional).
- Duplicate annotation vector for each dynamic MFCC.

Experiment: Vocabulary Pruning

- Web2131 Text corpus [Turnbull et al., 2006]
  - Collection of 2131 songs and accompanying expert song reviews mined from www.allmusic.com.
  - 315 word vocabulary.
  - Annotation vector is based on the presence or absence of a word in the review.
  - More noisy word-song relationships than CAL500.
- Experimental design
  - Merge vocabularies: 173+315=488 words.
  - Prune noisy words as we increase amount of sparsity in CCA.
- Hypothesis
  - Web2131 words will be pruned before CAL500 words.

Results: Vocabulary Pruning

Vocabulary size	488	249	203	149	103	50
# CAL500 words	173	118	101	85	65	39
# Web2131 words	315	131	102	64	38	11
%Web2131	**.64**	**.52**	**.50**	**.42**	**.36**	**.22**

Table: The fraction of noisy web-mined words in a vocabulary as vocabulary size is reduced: As the size shrinks sparse CCA prunes noisy words and the web-mined words are eliminated over higher quality CAL500 words.

Experiment: Vocabulary Selection for Music Retrieval

- $P(song|word)$ is modeled as a Gaussian mixture model.
- The system can annotate a novel song with words from its vocabulary or it can retrieve an ordered list of novel songs based on a keyword query.
- Evaluation metric for retrieval: Area under the ROC curve.

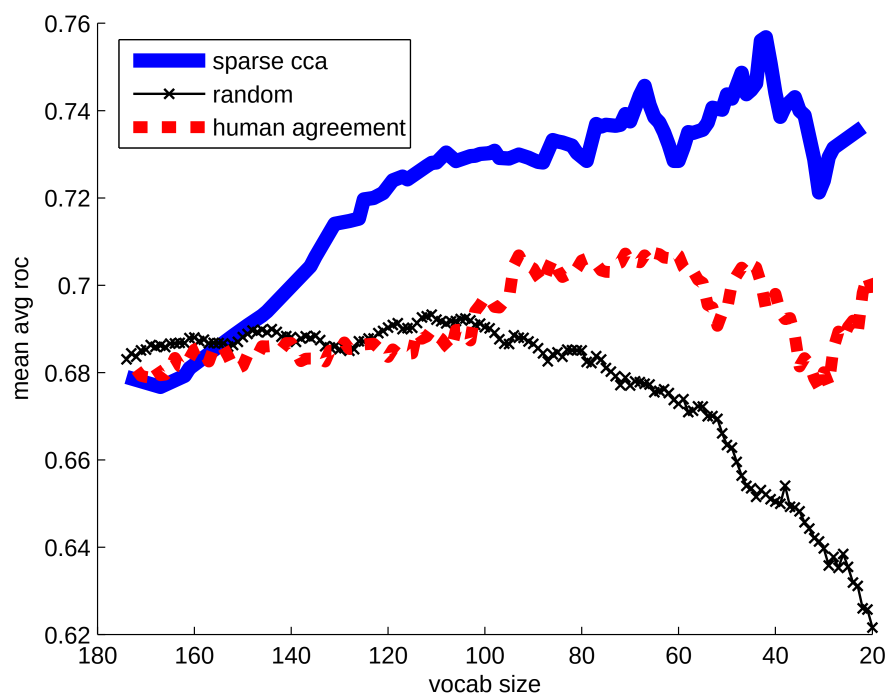


Figure: Comparison of vocabulary selection techniques: We compare vocabulary selection using human agreement, acoustic correlation, and a random baseline, as it effects retrieval performance.

Summary

- Constructing a meaningful vocabulary is the first step in building a content-based, natural-language search engine for music.
- Given a semantic representation and acoustic representation, sparse CCA can be used to find musically meaningful words.
  - semantic dimensions linearly correlated with audio features.
- Automatically pruning words is important when using noisy sources of semantic information.

References

Sriperumbudur, B. K., Torres, D., and Lanckriet, G. R. G. (2007).
Sparse eigen methods by d.c. programming.
In *ICML 2007*.

Tao, P. D. and An, L. T. H. (1998).
D.c. optimization algorithms for solving the trust region subproblem.
SIAM J. Optim., pages 476–505.

Turnbull, D., Barrington, L., and Lanckriet, G. R. G. (2006).
Modelling music and words using a multi-class naive bayes approach.
In *ISMIR 2006*.

Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. R. G. (2007).
Towards musical query-by-semantic description using the cal500 dataset.
In *SIGIR 2007*.

Questions

Thank You