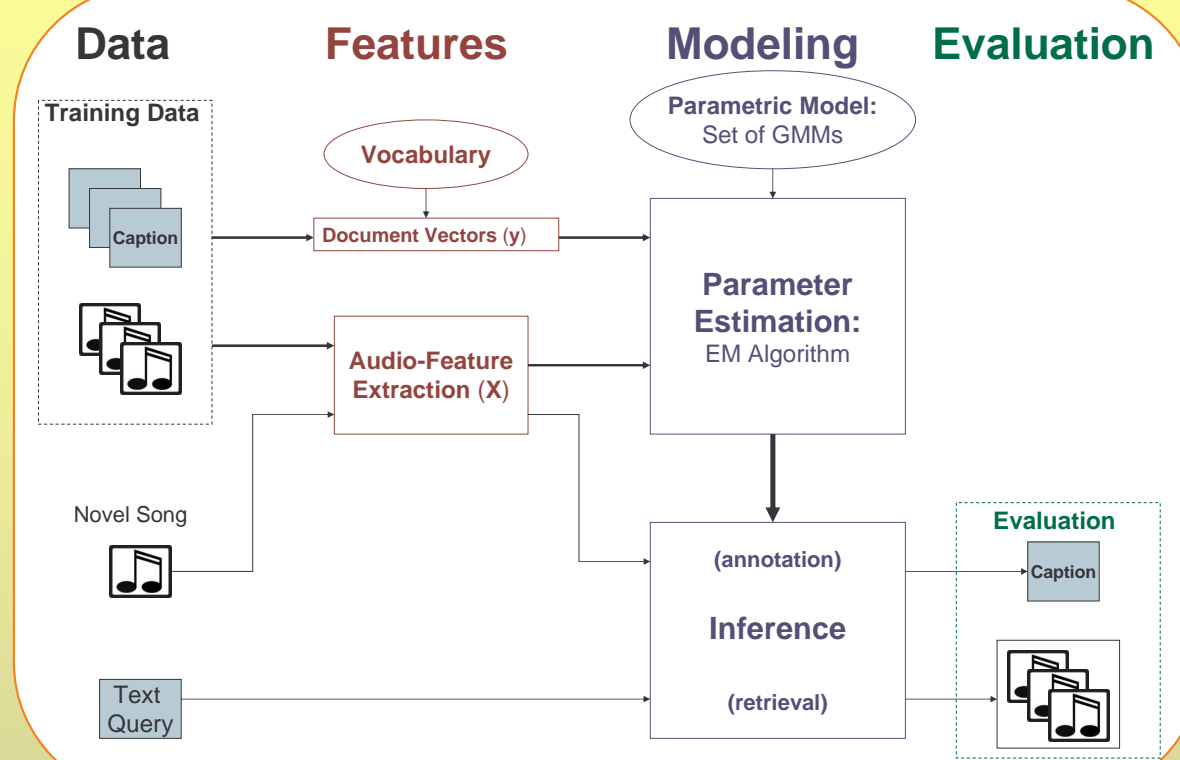


Motivation

Our goal is to design a **statistical system that learns a relationship between audio and words**.
Such a system can perform 2 tasks:

- Annotation:** Given an audio-content of a track, we can **'annotate'** the track with semantically meaningful words.
track → words
- Retrieval:** Given a text-based query, we can **'retrieve'** relevant tracks based on their audio content.
words → tracks



Audio & Text Features

Consider a **vocabulary** and a **heterogeneous data set of audio track-caption pairs**:

- Vocab - predefined set of words
- Track - bag of audio feature vectors, $X = \{x_1, \dots, x_T\}$
- Caption - binary document vector

Music:
DMFCC features summarize 3/4 second of audio content
Between 320-1920 feature vectors represent each song

SFX:
MFCC features plus 1st and 2nd time deltas

Annotation

Given 'word' distributions $P(x|i)$ and a **query track** $\{x_1, \dots, x_T\}$, we **annotate with word i^*** :

$$i^* = \arg \max_i P(i|x_1, \dots, x_T)$$

$$= \arg \max_i \frac{P(x_1, \dots, x_T|i)P(i)}{P(x_1, \dots, x_T)}$$

Naïve Bayes Assumption: x_i and x_j are conditionally independent, given i .

$$= \arg \max_i \prod_{t=1}^T P(x_t|i) \cdot P(i)$$

Assuming a uniform prior and taking the log, we have:

$$= \arg \max_i \sum_{t=1}^T \log P(x_t|i)$$

Using this equation, we **annotate the query track** with the top N words.

Retrieval

Given a query word q , ranking test songs by the **posterior probability** $P(x_1, \dots, x_T|q)$ results in almost the same ranking for all query words!

Length Bias
Longer tracks have proportionately lower likelihood

Song Bias
Many conditional word distributions $P(x|q)$ are similar to the generic *track* distribution $P(x)$

High probability (e.g. generic) tracks under $P(x)$ often have high probability under $P(x|q)$

Solution: Rank by **likelihood** $P(q|x_1, \dots, x_T)$ instead.
=> Normalize $P(x_1, \dots, x_T|q)$ by $P(x_1, \dots, x_T)$

Sound Semantics Model

For the i^{th} word in the vocabulary, estimate $P(x|i)$, a **'word' distribution over audio feature vector space**.
Model $P(x|i)$ with a Gaussian Mixture Model (GMM), estimated using Expectation Maximization

The training data for each word distribution is the set of all **feature vectors** from all tracks labeled with that word.

Multiple Instance Learning
training set includes some irrelevant feature vectors

Weakly Labeled Data
training set excludes some relevant feature vectors

Our **probabilistic model** is a set of 'word' distributions (GMMs)

Experimental Data

Music: 2131 popular western songs from last 60 years
Text: natural language song reviews from AMG Allmusic
Vocab: 317 'musically relevant' unigrams and bigrams

SFX: 1305 tracks from the BBC sound effects library
Text: track captions from the BBC library
Vocab: 348 words appearing 5 or more times

Annotation: annotate each test track with N words
Retrieval: rank order all test tracks given a query word

Results

		Annotation		Retrieval	
		Recall	Precision	mAP	mAROC
Music	Random	0.030	0.060	0.083	0.5
	Our Model	0.072	0.119	0.109	0.610
SFX	Random	0.013	0.011	0.05	0.5
	Our Model	0.243	0.220	0.201	0.793

Discussion

Music is **inherently subjective**
Different people use different words to describe the same song

We **learn and evaluate** using a noisy text corpus
Reviewers do not make explicit decisions about the individual words when reviewing a song.
"This song does **not rock**."
Mining the web may not suffice.

Sound effects results are comparable to the best results for content-based **image** annotation and retrieval.

References

Gimeno & Vasconcelos (2005). Formulating semantic image annotation as a supervised learning problem. *IEEE CVPR*.

Whitman & Ellis (2004). Automatic record reviews. *ISMIR*.

Slaney (2002). Semantic-audio retrieval. *IEEE ICASSP*.

McKinney & Brechtbart (2003). Features for audio and music classification. *ISMIR*.

AMG Allmusic guide. <http://www.allmusic.com>