Robust Music Identification, Detection, and Analysis

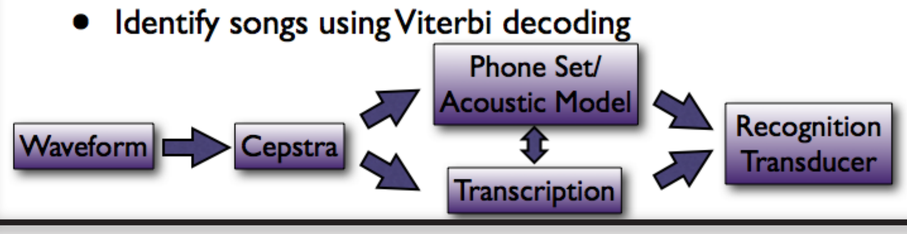
Mehryar Mohri[1,2], Pedro Moreno[2], Eugene Weinstein[1,2]
[1] Courant Institute of Mathematical Sciences; [2] Google Inc.

Introduction

- Music identification scenario: match a few seconds of audio to large song database
- Applications: music search, content monitoring, etc.
- Recording may be distorted or noisy; Only a short snippet of audio may be available
- Most past work based on hashing, e.g., [Haitsma et al. '01]
  - Match required between training and test features
- HMM system over music sound events: [Batlle et al. '02]
  - Automatically learn music "phonemes" describing songs

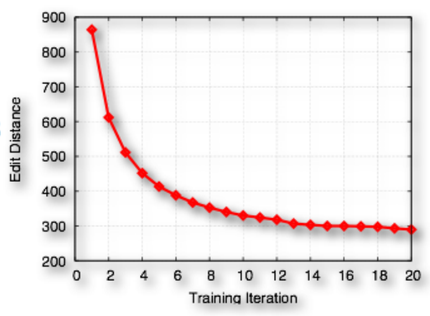
Overview

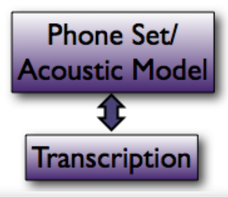
- Start with database of 15,000+ songs
- Compute MFCC features over audio
- Cluster song segments to get initial music phone set
- Learn phone set and train acoustic model for each phone
- Generate compact recognition transducer
- Identify songs using Viterbi decoding



Acoustic Model Training

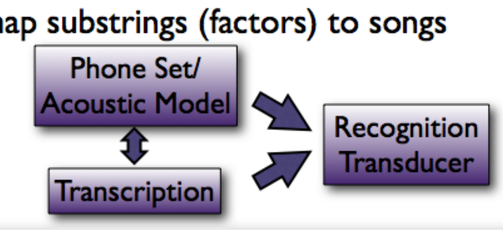
- Segment training audio based on spectral change
- Initialize model using k-means clustering over segments
- Iterative training, repeat:
  1. Run decoder with current model, get transcriptions
  2. Use transcription counts to train GMM for each phone
- Edit distance measures convergence





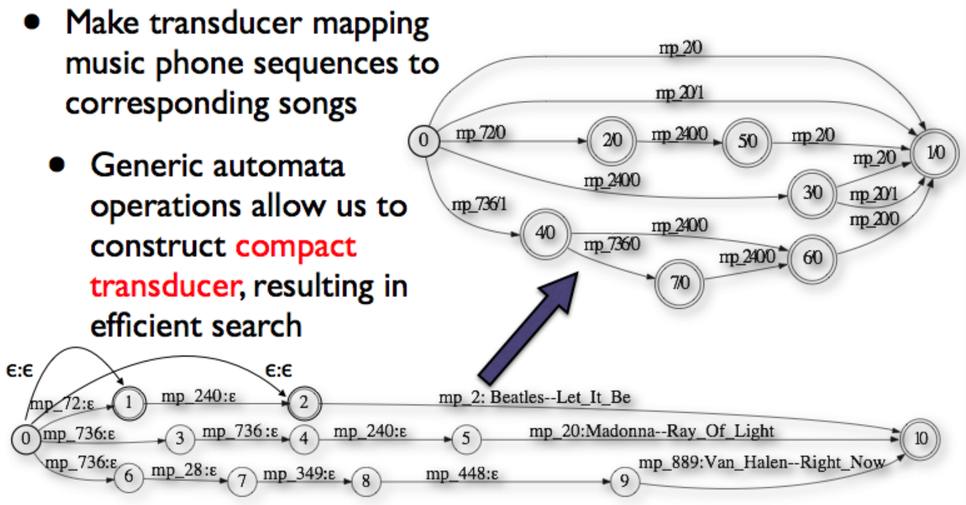
Finite-state Transducers

- Finite automata with input and output labels on transitions, possibly weighted
  - Widely used in speech and text processing, computational biology, etc.
- We view each song transcription as a string (music phones are the symbols)
- Goal: construct FST to map substrings (factors) to songs



Song Recognition

- Make transducer mapping music phone sequences to corresponding songs
  - Generic automata operations allow us to construct compact transducer, resulting in efficient search



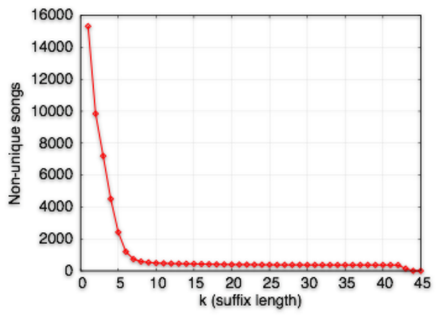
Experiments

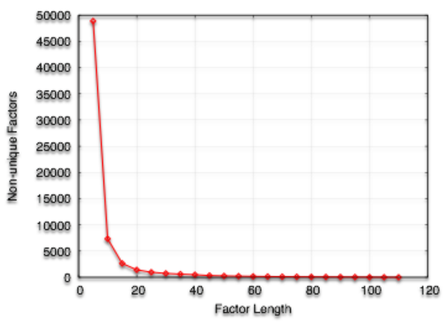
- Database: 15,455 songs in MP3 format
- Test set: 1,762 in-set and 1,856 out-of-set snippets
- Detection: distinguish in-set from out-of-set songs
  - Use Support Vector Machines to classify in-set/out-of set based on decoder scores

Condition	Identification Accuracy	Detection Accuracy
Clean	99.4%	96.9%
White noise @44.0dB SNR	98.5%	96.8%
White noise @24.8dB SNR	85.5%	94.5%
White noise @10.4dB SNR	39.0%	93.2%
White noise @5.9dB SNR	11.1%	93.5%
Speed up by 2%	96.0%	96.0%
Slow down by 2%	96.4%	96.4%
Speed up by 10%	43.2%	87.7%
Slow down by 10%	45.7%	85.8%
MP3 re-encode 64kbps	98.1%	96.6%
MP3 re-encode 32kbps	95.5%	95.3%

Suffix/Factor Uniqueness

- Number of "collisions" among song suffixes/factors drops off rapidly with increasing length





Summary

- "Phone" set for music ID can be learned automatically
  - Match audio without relying on directly comparing feature vectors
  - Robust to moderate noise and distortions
- We formulate music ID as a search/decoding problem
  - Use well-established techniques from speech and text processing (FSTs, GMMs) to construct effective system
  - Novel application of transducers allows efficient matching of song snippets