LARGE-SCALE CONTENT-BASED MATCHING OF MIDI AND AUDIO FILES

Alessandro Palmeira e Fábio Goródscy

IME-USP

04-08-2016

Conteúdo

Parte 1

Descrição

Objetivo

- Dado um arquivo MIDI, descobrir em um grande banco de áudio, versões de áudio que representem o arquivo

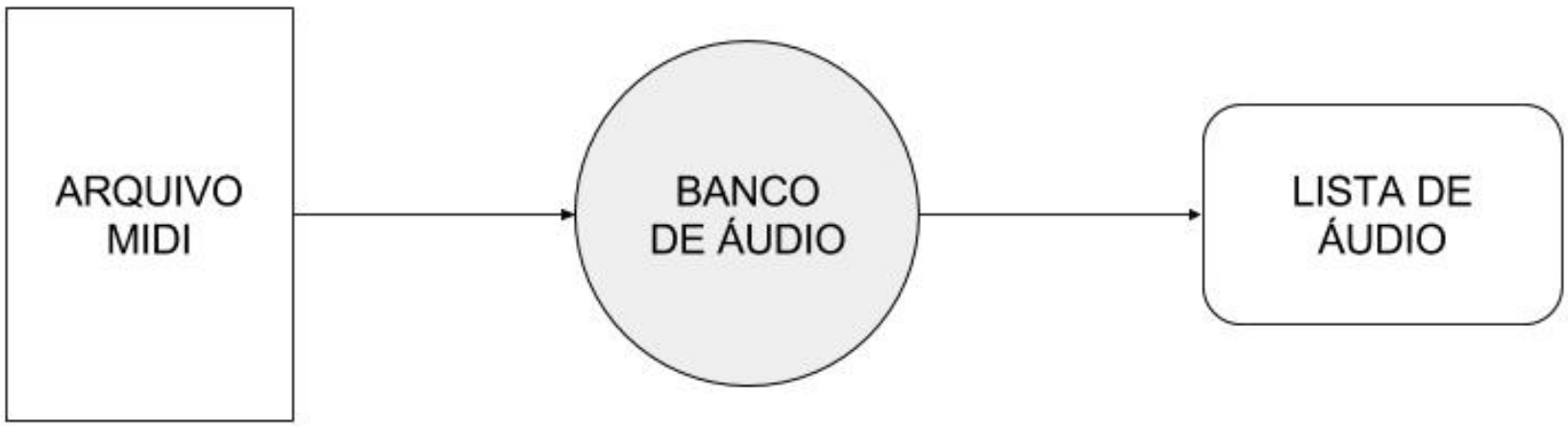

Figura 1: Esquema do sistema

Processo

1. Definir uma base de treinamento com pares MIDI/MP3

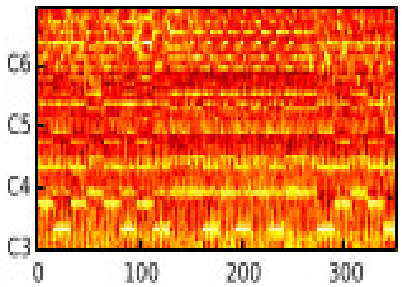

Figura 2: MP3

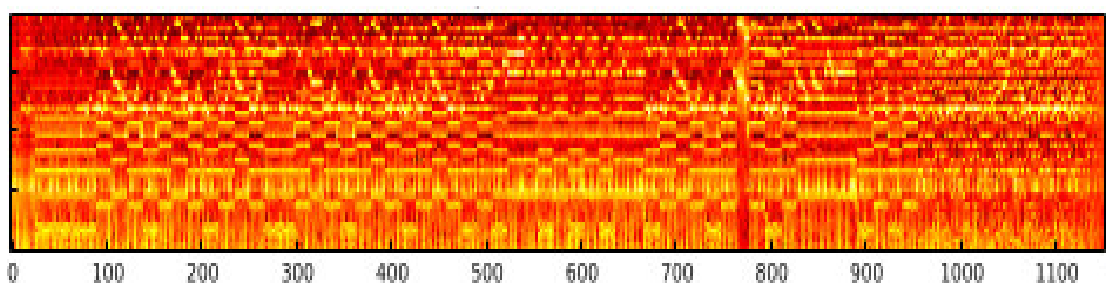

Figura 3: MIDI

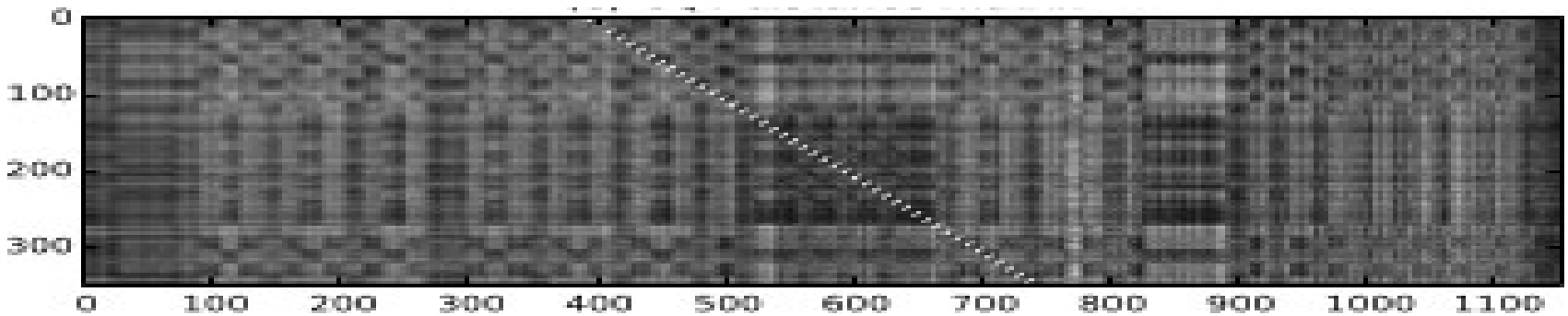

Figura 4: Alinhamento

Processo

2. Treinamento de rede neural capaz de aprender uma representação de vetores de espectro de Q constante (CQT) como vetores binários, baseado na base de treino

$$\begin{aligned}\mathcal{L} = & \frac{1}{|\mathcal{P}|} \sum_{(x,y)\in\mathcal{P}} \|f(x) - g(y)\|_2^2 \\ & - \frac{\alpha}{|\mathcal{N}|} \sum_{(x,y)\in\mathcal{N}} max(0, m - \|f(x) - g(y)\|_2^2)\end{aligned} \tag{1}$$

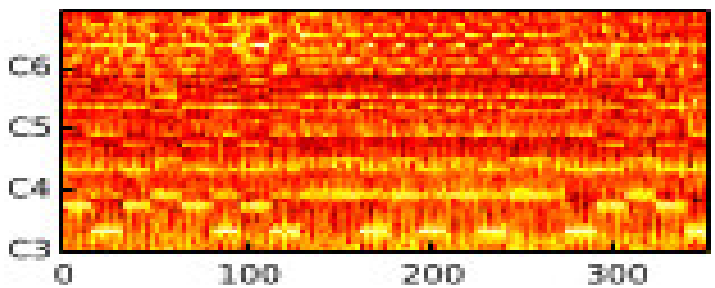

Figura 5: CQT

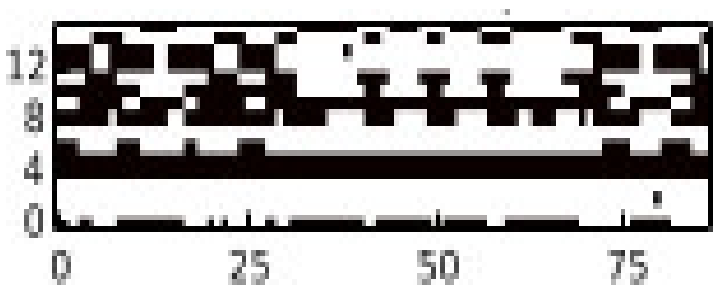

Figura 6: Hash

Processo

3. Utilizando a transformação aprendida pela rede neural calcula-se uma representação binária para cada mp3 do banco de áudio

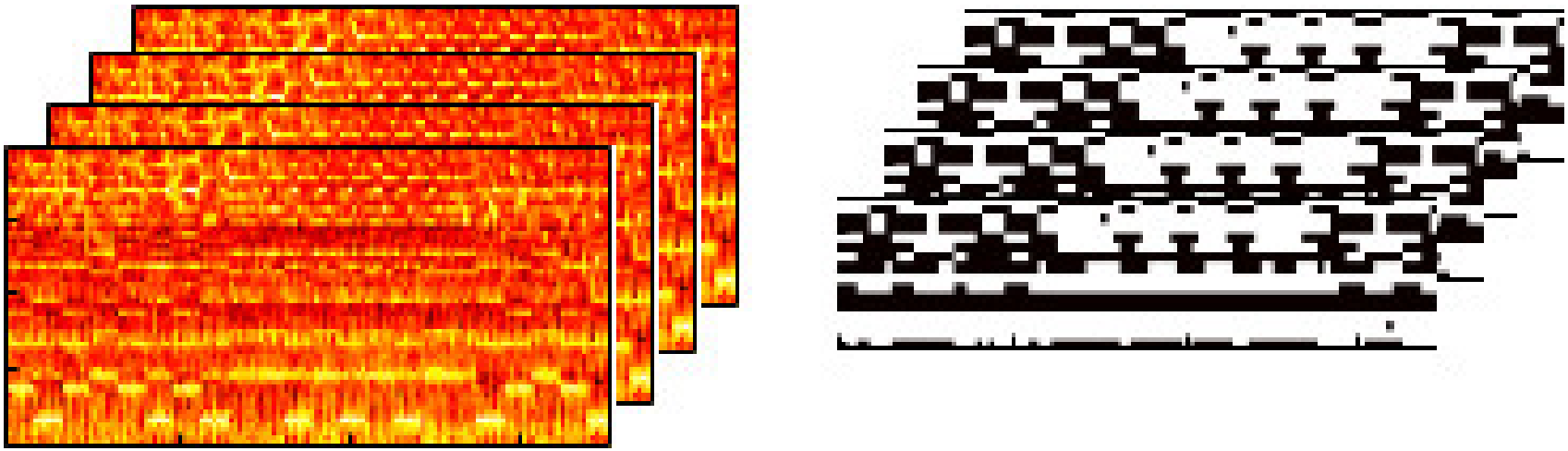

Figura 7: Representando banco como hash

Processo

4. Por fim, calcula-se a representação binária do arquivo MIDI da consulta e se compara (através de um DTW simplificado, onde os produtos internos são reduzidos a operações de ou exclusivo) com cada uma das representações do banco

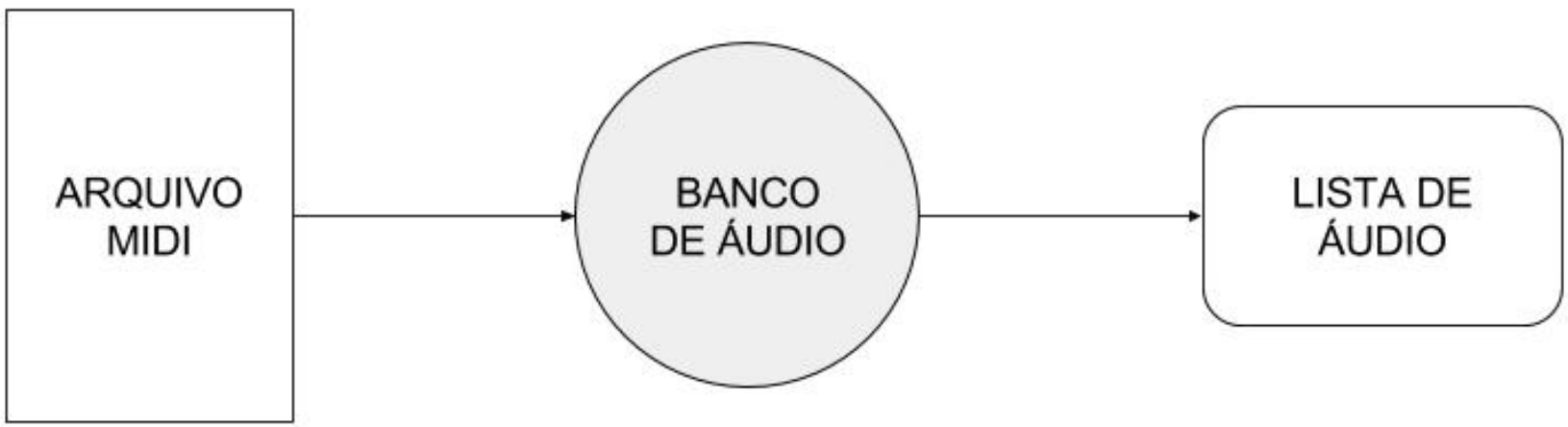

Figura 8: Arquivo MIDI é transformado em hash, comparado com banco

Parte 2

Datasets

MIDI Dataset

- 455,333 arquivos MIDI
  - 140,910 com MD5 checksum único
- Nem todos os arquivos tem informações de autor/título

J/Jerseygi.mid
V/VARIA18O.MID
Carpenters/WeveOnly.mid
2009 MIDI/handy_man1-D105.mid
G/Garotos Modernos - Bailanta De Fronteira.mid
Various Artists/REWINDNAS.MID
GoldenEarring/Twilight_Zone.mid

Tabela 1: Exemplos de arquivos

Clean MIDI Dataset

- 17,243 arquivos MIDI retirados do dataset inicial
  - 10,060 músicas diferentes
- Arquivos com informações autor/título
- Usado como entrada no treino da rede neural
- Também serve como ground truth para avaliar resultados da classificação

Million Song Dataset

- Dataset com features e metadata sobre 1 milhão músicas
- Além das informações disponibilizadas publicamente foram utilizados previews de 30 segundos de cada uma das músicas do msd.
  - Nos nossos testes utilizamos uma pequena porção disso, com 18,901 músicas.

Outros

- Também se somam aos datasets o CAL500, CAL10K e USPOP2002. Além das features públicas foram utilizados mp3 completos de cada música desses datasets
- Todos os datasets que possuem mp3 são utilizados como entrada do treino da rede neural, junto com a MIDI que a representa (a decisão da MIDI que representa um mp3 é baseada nos metadados de ambos os arquivos e em um alinhamento por dtw)

Parte 3

Resultados

Especificações

- Nos nossos testes ficamos com:
  - 2016 total de pares MP3/MIDI não alinhados
  - Base de treino para rede neural: 1225 pares alinhados
  - Base de validação da rede: 113 pares alinhados
  - 537 pares alinhados para ver os resultados
- Desses ficamos com 19 pares alinhados com um score do dtw suficientemente alto escolhido ao acaso para realizar o teste final.

Comparativo

Posição	Porcentagem de pares corretos na posição
1	15.2
10	41.6
100	62.8
1000	82.7
10000	95.9

Tabela 2: Resultados do teste obtido por Raffel et al.

Discução, comentários, dúvidas...

Obrigado!