LabROSA Research Overview

Dan Ellis

Laboratory for Recognition and Organization of Speech and Audio Dept. Electrical Eng., Columbia Univ., NY USA

dpwe@ee.columbia.edu

http://labrosa.ee.columbia.edu/

- I. Music
- 2. Environmental sound
- 3. Speech Enhancement



2014-06-12 - 1 /20



LabROSA Research Overview - Dan Ellis

LabROSA

Getting information from sound



I. Music Audio Analysis

Poliner & Ellis '06

- Trained classifiers for low-level information
 - notes, chords, beats, section boundaries
- E.g. Polyphonic transcription



feature agnosticneeds training data



2014-06-12 - 3/20

Million Song Dataset

Bertin-Mahieux McFee

- Industrial-scale database for music information research
- Many facets:
 - Echo Nest audio features
 + metadata
 - Echo Nest "taste profile" user-song-listen count
 - Second Hand Song covers
 - musiXmatch lyric BoW
 - last.fm tags
- Now with audio?
 - resolving artist / album / track / duration against what.cd



MIDI-to-MSD

Raffel

• Aligned MIDI to Audio is a nice transcription



• Can we find matches in large databases?

Singing ASR



Speech recognition adapted to singing

needs aligned data

• Extensive work to line up scraped

"acapellas" and full mix

• including jumps!

Project Rate (Hz): Selection Start: End • Length Audio Position: 1000 • • • Selection Start: 0 h 0 m 0 0.507 × 0 0 h 0 m 00.000 × 0	00							Elton	ohnYour_	Song-mix					
Project Rate (Hz): Selection Start: End End Length Audio Position: Book Snap To Solo to to mo 0.5007 x = Oh to mo 0.5007 x = 				- I 🛃	Ø L		R				<u></u> * p	<u>.</u>	·	<u>. +</u>	
Core A ‡ +0 Built-in Output ‡ Puilt-in Microph ‡ 2 (Ster ‡) 30 -15 0 75 30 45 100 1:15 1:30 1:45 2:00 2:15 2:30 2:45 3:00 × Max 40k 3.0k 3.0k <td< th=""><th></th><th></th><th></th><th>$\rho \leftrightarrow$</th><th>* •) ▼ -1</th><th>8 0</th><th><i>▶</i> -</th><th>18</th><th>o' 🛛 💥 🔤</th><th>5 🛸 -M-</th><th>nHn v</th><th><u>د</u></th><th>P</th><th>P 2</th><th>2</th></td<>				$ \rho \leftrightarrow $	* •) ▼ -1	8 0	<i>▶</i> -	18	o' 🛛 💥 🔤	5 🛸 -M-	nHn v	<u>د</u>	P	P 2	2
30 - 15 0 15 230 245 300 × Max + 40k - 25k - 200 2:15 2:30 2:45 3:00 × Max + 40k - 25k - 200 2:15 2:30 2:45 3:00 × Max + 40k - 25k - 20k - 25k - 2	Core A	(ب 🗧	Built-in Outp	ut 🛟 🔎	Built-in Microph	\$ 2	2 (Ster	\$							
X Mx 4.0k Stere, 16000Hz 3.5k 2.5k 2.0k 1.5k 1.0k 0.0k 0.0k 4.0k 3.5k 2.0k 1.5k 0.0k 0.0k 2.5k 0.0k 0.5k 0.0k <td< th=""><th>-30 -15</th><th></th><th>0</th><th>30</th><th>45 1:0</th><th>0</th><th>1:15</th><th>1:30</th><th>1:45</th><th>2:00</th><th>2:15</th><th>2:30</th><th>2:4</th><th>5</th><th>3:00</th></td<>	-30 -15		0	30	45 1:0	0	1:15	1:30	1:45	2:00	2:15	2:30	2:4	5	3:00
Arete, noormal 3.5k - Wret Solo Job 2.5k - Job 1.5k - Job 0.5k -	X Mix	▼ 4.0k		化化化化化				A.						.41	1 1 3
Mute Solo 3.0k+	32-bit float	- 3.5k						- Aler			de des Se		1.1		
Project Rate (Hz): Selection Start: End End Log Do h 00 m 06.113 st Do h 00 m 09.607 st Do h 00 m 09.607 st 	Mute Solo	3.0k-	1211月1	经合作化计	P. P. M. N. WILL		the file	Astellan	West Street Street	1-40 4 24	a defe dana Aliantia (dente			Lar +	a line la
Project Rate (Hz): Selection Start: End @ Length Audio Position: 1600 @ Snap To 00 h 00 m 06.113 sy 00 h 00 m 09.607 sy 00 h 00 m 00.000 sy	<u> </u>	2.5k	pate aparte acente de la contrate					Tal Local	an spran a start						1 L
Project Rate (Hz): Selection Start: End @ Length Audio Position: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: 1600 Image: Selection Start: Image: Selection Start: Image: Selection Start: <td< th=""><th><u> </u></th><th>2.0k-</th><th>Landar Harbert</th><th>教教教教会</th><th></th><th></th><th>5.1.11</th><th>in a surry</th><th></th><th>A TANK</th><th>with the</th><th></th><th></th><th></th><th></th></td<>	<u> </u>	2.0k-	Landar Harbert	教教教教会			5.1.11	in a surry		A TANK	with the				
I.0k- 0.5k- 0.0k 3.0k- 2.5k- 2.0k- 1.5k- 1.0k- 0.0k 2.0k- 1.5k- 0.0k 0.0k Project Rate (Hz): Selection Start: End <a a="" length<=""> Audio Position: 00 h 00 m 09.607 s 16000 Selection Start: 0 h 00 m 09.607 s 16000 0 h 00 m 06.113 s 00 h 00 m 09.607 s		1.5k		1110		1		Shur-		1.111	設設での				Bay.
0.0k 0.0k 4.0k 3.5k 3.0k 2.5k 2.0k 1.5k 1.5k 0.0k 0.0k 0.0k Project Rate (Hz): Selection Start: End 16000 Image: Selection Start: End Length Audio Position: 100 h 00 m 06.113 sm 00 h 00 m 09.607 sm 00 h 00 m 00.000 sm		1.0k-		We want the set		1	sense trate	THE A						14.	100 - 100 -
4.0k 3.5k 3.0k 2.5k 2.0k 1.5k 1.0k 0.0k 9 Project Rate (Hz): Selection Start: End © Length Audio Position: 16000 © nap To 00 h 00 m 06.113 sv 00 h 00 m 00.000 sv		0.5k 0.0k	and the second s	al a la d'ite			and the second second				165				
3.5k- 3.0k- 2.5k- 2.0k- 1.5k- 1.5k- 0.0k		4.0k				120		1500 151 1	80.082		調測機器		10.0100	福田	ALL N
3.0k- 2.5k- 2.0k- 1.5k- 1.0k- 0.5k- 0.0k		3.5k	- 88.01					1		日間周囲					NI.
2.5k- 2.0k- 1.5k- 1.0k- 0.5k- 0.0k Project Rate (Hz): Selection Start: End Length Audio Position: 16000 Find Length Audio Position: 1000 0 Click and drag to select audio		3.0k-												12 h	\$11
2.0k- 1.5k- 1.0k- 0.0k 0.0k Project Rate (Hz): Selection Start: End < Length Audio Position: 16000 Snap To 00 h 00 m 06.113 sv 00 h 00 m 09.607 sv 00 h 00 m 00.000 sv		2.5k	- 884	影響的構成。				1351			調整 清核			R.F.	
1.5k- 1.0k- 0.5k- 0.0k Project Rate (Hz): Selection Start: End Length Audio Position: 16000 Click and drag to select audio		2.0k-			王家居的部族						調整開作				311
I.0k- 0.5k- 0.0k I.0k- 0.5k- 0.0k Project Rate (Hz): Selection Start: End <a length<="" td=""> Audio Position: 16000 Image: Ima		1.5k		SEA NE	前来(時)。開告			1410			調査 周上	- 情報書		100	21
O.Sk- 0.0k O.Sk- 0.0k Project Rate (Hz): Selection Start: End <a length<="" td=""> Audio Position: 16000 Image: Selection Start: Do h 00 m 06.113 sy 00 h 00 m 09.607 sy 00 h 00 m 00.000 sy		1.0k-			长马伯恩器的			the A	發展	1451	1. 1. 1.		1218		211
O.0k Image: Selection Start: End Length Audio Position: 16000 Image: Selection Start: 00 h 00 m 09.607 s 00 h 00 m 00.000 s		0.5k	11.5	影出我的意	計劃性成為的			index .	建築		WH HE		1	NO.	MR.
Project Rate (Hz): Selection Start: End <a height="list-based list-based list-base</th> <th></th> <th>0.0k</th> <th>400 den 23.</th> <th>mi and the first</th> <th>Print Pro Line Print Pri</th> <th>Calle Ann Col</th> <th>and the states for</th> <th>N THE R</th> <th></th> <th>ne - And and it</th> <th>the second the</th> <th>s an in the</th> <th>a con an stan</th> <th>Marty Care</th> <th></th>		0.0k	400 den 23.	mi and the first	Print Pro Line Print Pri	Calle Ann Col	and the states for	N THE R		ne - And and it	the second the	s an in the	a con an stan	Marty Care	
Project Rate (Hz): Selection Start: End <a length<="" th=""> Audio Position: 16000 Image: Shap To 00 h 00 m 06.113 s m 00 h 00 m 09.607 s m 00 h 00 m 00.000 s m															
16000 Snap To 00 h 00 m 06.113 s 00 h 00 m 09.607 s 00 h 00 m 00.000 s	Project Rate	(Hz):	Se	lection Start:) End	💿 Leng	th	Audio	Position:						
Click and drag to select audio	16000 🛟		Snap To 0	0 h 00 m 06.1	113 s - 00 h 0	0 m 09.	.607 s •	00 h (00 m 00.00	0 s •					
	Click and drag t	to select	audio												

Block Structure RPCA

Papadopoulos

- RPCA separates vocals and background based on low rank optimization
 - single trade-off parameter
 - adjust based on higher-level musical features?



Ordinal LDA Segmentation

- Low-rank decomposition of skewed self-similarity to identify repeats
- Learned weighting of multiple factors
 - to segment
 - Linear
 Discriminant
 Analysis
 between
 adjacent
 segments



McFee

2. Environmental Sound

- Extracting useful information from soundtracks
- e.g.TRECVID Multimedia Event Detection (MED)
 - "Making a Sandwich", "Getting a Vehicle Unstuck"
 - 100 examples, find matches in 100k videos
 - manual annotations for ~ 10 h



E009 Getting a Vehicle Unstuck

LabROSA Research Overview - Dan Ellis

Foreground Event Recognition

Cotton, Ellis, Loui 'I I



• Transients = foreground events?

- Onset detector finds energy bursts
 - best SNR
- PCA basis to
 - represent each
 - 300 ms x auditory freq

• "bag of transients"

NMF Transient Features

Decompose spectrograms into templates + activation

Smaragdis & Brown '03 Abdallah & Plumbley '04 Virtanen '07 Cotton & Ellis' 11



LabROSA Research Overview - Dan Ellis

2014-06-12 - 11/20

Background Retrieval

McDermott et al. '09 Ellis, Zheng, McDermott '11

Classify soundtracks by statistics of ambience
E.g. Texture features



- Subband distributions
- Envelope cross-corrs



LabROSA Research Overview - Dan Ellis

2014-06-12 - 12/20

Auditory Model Features

Lyon et al. 2010 Lee & Ellis 2012 Cotton & Ellis 2013

- Subband Autocorrelation PCA
 - Simplified version of autocorrelogram
 - IOx faster than Lyon original
- Capture fine time structure in multiple bands
 information lost in MFCCs



Subband Autocorrelation

 Autocorrelation stabilizes fine time structure





- 25 ms window, lags up to 25 ms
- calculated every 10 ms
- normalized to max (zero lag)

Retrieval Examples

High precision for in-domain top hits



LabROSA Research Overview - Dan Ellis

2014-06-12 - 15/20

3. Speech Enhancement

- Noisy speech scenarios
 - Ambient recording (background noise)
 - Communication channel (processing distortion)



RPCA Enhancement

Chen, McFee & Ellis 'I 4

- Decompose spectrogram into sparse + low-rank
- Sparse activation H of dictionary W

 $\min_{H,L,S} \lambda_H \|H\|_1 + \lambda_L \|L\|_* + \lambda_S \|S\|_1$

 $+\mathcal{I}_+(H)$

- s.t. Y = WH + L + S
 - ASR benefits:

	С	S	D	I
Orig	6.8	10.6	82.6	0.7
RPCA	10.8	36.5	52.7	0.5
wie+RPCA	10.4	40.1	49.5	2.1



Classification Pitch Tracker

Lee & Ellis 'I 2

 SAcC: MLP trained on noisy speech with ground-truth pitch track targets



 Large benefits for in-domain noisy speech



Pitch-Normalized Enhancement

• Use noise-robust pitch tracker for enhancement?

- Normalize
 voice pitch
- Fixed-pitch enhancement
- Reimpose pitch



Summary

Music

- transcription, segmentation, ...
- alignment for ground truth

Soundtracks

• foreground events, background ambience

Noisy Speech

- classification pitch tracking
- spectrogram enhancement