NSF-CAREER: The Listening Machine
IIS-0238301
Annual Report 2007

Daniel P.W. Ellis
Department of Electrical Engineering,
Columbia University
500 W. 120th Steet, New York NY 10027, USA

dpwe@ee.columbia.edu

Feb 12, 2008

1 Activities & Findings

1.1 Research Activities

We continued our investigations into extracting high-level information from audio, focusing on the domains of soundtracks of environmental/consumer recordings, music audio, and marine mammal sounds.

1.1.1 Environmental Sound Recognition

We have continued our research into associating words with the soundtracks of recordings of natural environments. We have been working with a database of 1400 "consumer videos" (collected by our collaborators at Kodak) as well as with similar amateur videos downloaded from YouTube. Based on a provisional lexicon of 25 terms that consumers might use as search terms ("music", "birthday", "beach"), we have experimented with a range of feature representations and classification schemes for automatically tagging new videos. This approach can be surprisingly effective, even for classes (e.g. “museum”) whose acoustic correlates are not necessarily obvious [1].

As part of this work, we have noticed that music is a particularly useful special case, so we have developed special-purpose features to support music detection. Our approach has been to look for two, complementary, characteristics of music - pitches and rhythm [6]. Pitches are detected via the stability of periodicity peaks in autocorrelation averaged over a window of 100ms or more, and rhythm is detected via our beat-detector which looks for peaks in the autocorrelation of the overall energy envelope that arise from rhythmic ‘events’ in the music [3].

In the domain of long-duration personal audio recordings, we presented our work on using audio hashes to efficiently identify and find repeating sounds in long recordings [7]. We have also collaborated with colleagues in Dublin on segmenting long ‘lifelog’ recordings by combining both audio- and video-based innovation measures [2].

1.1.2 Music audio

In September 2007, we began a separate, NSF-funded project on Data-Driven Music Understanding (IIS-07-13334) which is based upon the music audio understanding work performed in this project, and henceforth our music work will be reported to that project. Prior to commencing that project, however, we continued to work on our cover-song detection system, which matches alternative performances of the same underlying song by representing each piece as a beat-synchronous array of instrumentation-invariant 12-dimensional chroma features. Matches are found by global cross-correlation of these beat-chroma matrices (efficiently implemented via the FFT), which finds the single best alignment point between different pieces. While this is unlikely to give perfect alignment over the whole track, finding even a significant subsection that correlates closely almost surely indicates a cover version [4].

1.1.3 Marine mammals

We continue our work on developing tools useful to biologists working with marine mammal calls. It turns out that the deceptively simple task of detecting when a dolphin call is present very often involves such a poor signal-to-noise ratio that conventional methods, which apply a simple threshold to the total energy of the signal, will make many mistakes. Instead, we are systematically investigating a range of different features and classifiers. Our recent paper analyzes the shortfalls of conventional pitch recognizers (designed for speech) [5].

1.2 Educational activities

As the core of the research in my lab, this project continues to feed input into various classroom activities. For the undergraduate/masters Digital Signal Processing course, the research material provides ideas and data for the small individual class projects completed by each student. This year, we had several projects on separating sounds (e.g. voices from noise or music) that draw on the general ideas of this project. For the graduate level Speech and Audio Processing and Recognition class, most of the more substantial individual projects fall squarely into the area of this course (which provides a number of project idea suggestions). In addition, several weeks' worth of the course content draws on material from this project, including music analysis, signal separation, and audio archive indexing. Also this year I taught a senior undergraduate class on music signal processing that employed some of the music analysis work performed under this project.

1.3 Findings

The findings from the research projects listed above include:

- Consumer audio-video recordings of the kind that are becoming increasingly common with the latest digital cameras may be effectively categorized on the basis of their soundtracks alone, achieving average precisions as high as 0.7 for user-relevant tags such as “has music”, yet still able to achieve useful average precisions of around 0.3 for tags with less obvious connection to the soundtrack such as “sports”, “park”, or “crowd”. Combining this audio-based information with video features provides significant improvements over either modality alone [1].

- Implementing special-purpose detectors for audio categories such as music can give significant improvements. By combining detectors specially produced to detect the rhythmic and tonal features that are characteristic of music, we are able to improve average precision to 0.95, compared to 0.8 when using conventional features, on a database of around 2000 everyday video soundtracks collected at random from YouTube [6].

- Fingerprinting techniques devised for recognizing music tracks in noisy environments can also be used to efficiently search across very large everyday sound databases and identify all recurring sounds – provided they show high spectral and temporal consistency in their repetitions, such as cell-phone ringtones, or the theme tunes of broadcast media programs [7].

- For long-duration “lifelog” audio-video recordings, change-based segmentation of the audio is a valuable adjunct to video segmentation based on visual features when attempting to decompose the recording into distinct locations, activities, etc. [2].

- A beat-synchronous representation of the musical chroma at each point in a music recording is an effective representation for matching “cover” performances of the same song by different artists. Global cross-correlation can identify matches effectively and quite quickly [4].

- Conventional pitch tracking algorithms, developed and optimized for speech in low-noise conditions, fare poorly on recordings of dolphin calls, which both exhibit a much wider range of pitches, and are typically embedded in high noise backgrounds. This unusual regime for pitch tracking requires a more robust approach which we are currently investigating [5].

2 Training and development

This year, the project has provided support for graduate students Keansub Lee, Michael Mandel, Graham Poliner, and Xanadu Halkias, who performed the research described in the earlier sections. Each of these students is working towards a Ph.D., and their research experience in this project is an essential part of their development into full members of the research community.

3 Outreach

This year saw presentations by the PI and students from LabROSA at IEEE ICASSP (Hawai'i), IEEE Mohonk Workshop on Audio Applications, Aalborg/DTU Intelligent Sound Workshop (Denmark), Radcliffe Institute for Advanced Studies, and Kodak Research Labs (Rochester).

The PI completed co-editing a special issue of IEEE Tr. Audio, Speech, and Language, on the topic of Music Information Retrieval which appeared in February 2008.

We are organizing a third workshop on Statistical and Perceptual Audition SAPA-2008, to be held in September 2008.

4 Contributions

4.1 Contributions to the principal discipline

Our approach to classifying soundtracks into broad categories based on user tags, and in particular using online resources such as YouTube as data sources, represents an important advance in this area.

4.2 Contributions to sister disciplines

The work in music processing has interesting potential to help with musicological work that involves managing large databases.

The marine mammal sound analysis is principally motivated by the needs of biologists studying the behavior and communication of dolphins.

4.3 Contributions to human resources

The project has provided direct (partial) support for four graduate students, and has enabled the PI to participate in research supervision of several other students.

4.4 Contributions to educational infrastructure

The project has supported the continuing development of classroom and practical materials which we make freely available on our web site. These materials have been found valuable by many academics at multiple institutions worldwide.

4.5 Contributions to wider aspects of public welfare

Much of our work points directly to applications with the potential for wide impact in society at large. Our work on personal audio archive management deals with a kind of personal data record which we see as becoming more and more widespread, just as a very wide community has over the past few years begun to accumulate a ‘history’ of email communications, with the concomitant demand for archive access tools.

5 Resources made available

The following list recaps the online resources related to this project made available at the PI's web site:

- MEAPsoft - software for analyzing, visualizing, and reordering music audio recordings: `http://www.meapsoft.org/`
- Class materials (slides, assignments, practicals, demonstrations) for Digital Signal Processing: `http://www.ee.columbia.edu/~dpwe/e4810/`
- Class materials (slides, assignments, practicals, demonstrations) for Speech and Audio Processing and Recognition: `http://www.ee.columbia.edu/~dpwe/e6820/`
- Class notes and self-guided practical for the short course in Music Content Analysis by Machine Learning: `http://www.ee.columbia.edu/~dpwe/muscontent/`
- Focused collection of Sound Examples for use in student projects: `http://www.ee.columbia.edu/~dpwe/sounds/`
- Matlab examples of common audio processing algorithms: `http://www.ee.columbia.edu/~dpwe/resources/matlab/`
- Proceedings of the 2006 workshop on Statistical and Perceptual audition: `http://www.sapa2006.org/`

6 Publications

See references [1, 6, 3, 7, 2, 4, 5].

References

[1] Shih-Fu Chang, Dan Ellis, Wei Jiang, Keansub Lee, Akira Yanagawa, Alexander C. Loui, and Jiebo Luo. Large-scale multimodal semantic concept detection for consumer video. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 255–264, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-778-0. doi: http://doi.acm.org/10.1145/1290082.1290118. URL `http://www.ee.columbia.edu/~dpwe/pubs/ChangEtc07-consumer.pdf`.

[2] A.R. Doherty, A.F Smeaton, K. Lee, and D. P. W. Ellis. Multimodal segmentation of lifelog data. In *Proc. RIAO meeting*, 2007.

[3] D. P. W. Ellis. Beat tracking by dynamic programming. *J. New Music Research*, 36(1):51–60, March 2007. doi: 10.1080/09298210701653344. URL `http://www.ee.columbia.edu/~dpwe/pubs/Ellis07-beattrack.pdf`. Special Issue on Tempo and Beat Extraction, to appear.

[4] D. P. W. Ellis and G. Poliner. Identifying cover songs with chroma features and dynamic programming beat tracking. In *Proc. ICASSP*, pages IV–1429–1432, Hawai'i, 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/EllisP07-coversongs.pdf`.

[5] X. Halkias and D. P. W. Ellis. A comparison of pitch extraction methodologies for dolphin vocalization. *Canadian Acoustics*, page (to appear), 2008.

[6] K. Lee and D. P. W. Ellis. Detecting music in ambient audio by long-window autocorrelation. In *Proc. ICASSP*, page (to appear), 2008.

[7] J. Ogle and D. P. W. Ellis. Fingerprinting to identify repeated sound events in long-duration personal audio recordings. In *Proc. ICASSP*, pages I–233–236, Hawai'i, 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/OgleE07-pershash.pdf`.