NSF-CAREER: The Listening Machine
Annual Report 2006

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

dpwe@ee.columbia.edu

Mar 13, 2007

1 Activities & Findings

1.1 Research Activities

In 2006, we continued our investigations in machine understanding of complex audio signals in the domains of environmental 'personal audio' recordings, for music audio, and for marine mammal sounds.

1.1.1 Environmental Sound Separation

The problem of isolating and identifying individual sources in environmental recordings has been pursued from a number of directions. Ron Weiss has continued work on separation sound mixtures based on the constraints provided by learned models of source signal behavior (specifically speech); this work is now feeding into our new NSF supported collaboration on Separating Speech from Speech Noise (IIS-05-35168). We published a description of this work at ICASSP [5], as well as work on estimating time-frequency masks based on a trained "Relevance Vector Machine" classifier which we presented at the SAPA workshop [15].

Spatial information can be very helpful in separating sounds, particularly when multiple channels are available. We have been looking at recovering information on source location from two-channel recordings such as might be recorded from a 'dummy head' microphone. Although this is a well-studied problem, the presence of reverberation (where reflections from walls can appear as a large number of additional 'ghost' sources) generally defeats current techniques. We used statistical modeling to make a precise fit to the between-channel characteristics that arise from reverberation, enabling accurate inference of the position of multiple, overlapping sound sources, and resulting in a time-frequency mask to recover the signal for each source. This work was presented at the SAPA workshop [10] and at the prestigious Neural Information Processing Systems conference [11].

1.1.2 Personal Audio Organization

In addition to publishing an overview article on our earlier work in segmenting and classifying 'personal audio' (sound archives as might be collected by body-worn recorders) [8], we have focused specifically on identifying voice segments, as these generally indicate interesting events or occasions. Because the noise level is often very high in these recordings, we could not rely on conventional spectral models to identify speech. Instead, we addressed the problem of picking out the voice pitch track, which, perceptually, is often the most robust cue to vocal activity (for instance, consider hearing people talk in the next room, even though what they say is completely unintelligible). Our work on highly-noise-robust pitch tracking was able to detect speech in our recordings; in working to reduce false alarms from background hums such as air conditioning, we further developed an online adaptation scheme that has promising applications in separating speech from more stationary background sounds like music [9].

At a different level, we are currently looking at the problem of associating words with environmental audio recordings. In order to come up with the most useful words, as well as to collect essentially unlimited amounts of training data, we have been looking at the "tags" associated with amateur videos uploaded to the YouTube web site. By using their search feature to return videos associated with particular tags, we can collect large training and test sets; this work is currently being prepared for publication.

Another approach to organizing these long recordings is to look for all repeating sound events – for instance, a telephone ring, or the theme music to a radio program, or other daily sounds. In general, this involves an extremely large exhaustive comparison of all frames against all others, but we looked at using audio

fingerprint hashes to speed this up. Using an approach developed for identifying music in high noise conditions, we encode each 6 sec chunk of the recording into a small number of hashes from a finite (24 bit word) set. We can then quickly identify segments that share multiple hashes, and do a more careful comparison of those pairs. This work will be presented at ICASSP'07 [12].

1.1.3 Music audio

We have continued our successful innovations in applying techniques from machine listening to music audio. In our innovative data-driven approach, transcription (i.e. recovering the notes played from the acoustic recording) is cast as a conventional classification problem, based on a training set of acoustic feature vectors and associated target labels – obtained either from synthesized music (where the notes at each instant are known in advance e.g. from MIDI streams), or from multitrack originals (where notes can be automatically detected in solo tracks, then used as ground truth for the final mixdown). Papers published this year describe this approach applied both to general melody detection [3], and to polyphonic transcription (detecting all simultaneous notes) for piano music only [13]. We also had a paper accepted describing and interpreting the international melody extraction evaluation that we both participated in and ran in 2005 and 2006 [14].

This year we also participated in an evaluation for detecting "cover songs" – the same musical piece performed by different artists with different instrumentation and possibly a very different musical style. The essence of our system was to use beat tracking to obtain a representation that was invariant to variations in tempo, and a spectral feature that collapsed the entire frequency range onto 12 bins corresponding to the twelve distinct semitone chroma of the musical octave [7, 6]. Our system came top in the 2006 MIREX Cover Song evaluation, identifying more than twice as many cover versions as the next best submission [2]. Further experiments and evaluation have been described in a paper to appear at ICASSP 2007 [4]. Our interest in this is not specifically the narrow application in finding cover songs, but in developing a representation that captures melodic/harmonic content independent of instrumentation. Our current work is looking at clustering harmonically-related fragments within the entire output of an artist, or a larger music collection. Our motivation is to identify possible underlying common large-scale features within an entire music cannon.

Our broadest goal in the music work is to develop applications in similarity-based browsing and music recommendation based solely on the audio signal. We are continuing our survey work on music listening habits with local high-school

student Jeff Bauer. Also, Adam Berenzweig graduated with his Ph.D. after completing his thesis on similarity browsing, which including several new results on the current problem of 'hubs' (particular songs that are rated as similar to an unreasonably large proportion of the database) [1].

Finally, our experimental collaborative projects class conducted with Columbia's Computer Music Center bore fruit in the form of a software package, MEAPsoft, for analyzing, visualizing, and rearranging music recordings [16]. This software, which is freely available from our web site, has been picked up by several experimental composers and enthusiasts. However, we see great educational potential for its visualization capabilities, and we are developing a new version aimed at that application.

1.1.4 Marine mammals

Our work in analyzing marine mammal sound analysis, started last year, has continued. We have been working with prominent dolphin biologist Dr. Diana Reiss to develop automatic techniques for extracting dolphin whistles from high-noise underwater recordings. Describing these calls from large archives is very valuable for developing and verifying theories of animal communication; currently, such work relies on manual annotation and this is typically based on corpora of a few hundred examples at best. Our goal is to develop an automatic system, which we can verify against such existing annotated data, but then use to analyze recordings several orders of magnitude longer in duration.

1.2 Educational activities

This project supported developments to the three regular classes I taught this year:

- **Digital Signal Processing** (ELEN E4810), our foundation senior/masters level signal processing class, involving 50-60 students. Research results formed the basis for in-class demos and several of the semester projects.
- **Speech and Audio Processing and Recognition** (ELEN E6820), my graduate research-oriented class includes modules on signal separation, music signal analysis, and lrage audio archives, each drawing richly on this project. The student projects that count for half the grade are frequently based on ongoing research, and this year one of them led to a major publication [12].

- **Music Engineering Art Projects** (MEAP), a weekly interdisciplinary seminar involving engineering, composition, and other students. This year's main focus was the development and release of MEAPsoft [16] (described above). One of the motivations behind MEAPsoft is to investigate how our music audio analysis algorithms can be useful in creative projects, but it is also uncovering unexpected educational possibilities – to give people new insights into the structure and nature of their music, and as a possible way to make science and engineering accessible and relevant to middle- and high-school students. We are in discussion with the NSF GK12 school outreach project at Columbia Engineering to investigate this further.

I also gave several invited talks, including an overview of the personal audio analysis given at Microsoft in July, a tutorial on sound separation given at the AES meeting in Paris in May, and an overview of the music analysis given at a meeting of the Danish Intelligent Sound project in May.

1.3 Findings

The findings from the research projects listed above include:

- Speech signals can be separated with good perceived quality via a combination of identifying time-frequency cells dominated by a particular source, then fitting pre-trained models of common speech sounds to the partially-observed spectra. The "good" time-frequency cells can be identified with separate models trained to recognize the local characteristics of unmasked speech components.

- When two or more channels are available, between-channel differences provide strong and independent information concerning which cells belong to which source. Between-channel time differences, while often noisy and sometimes ambiguous, can be integrated within a rigorous probabilistic framework to detect the directions (time differences) associated with different sources, and the time-frequency cells most strongly dominated by those sources.

- Subband autocorrelation proved able to detect voices via their pitch even in recordings so noisy that the speech itself is no longer intelligible. Common decoys, such as the steady hum of machinery or even music, can be successfully rejected based on their locally stationary period, in contrast to the constantly-changing voice pitch.

- To associate specific words with recorded multimedia content, consumer media sites such as YouTube can be a valuable source of training data. Classifiers trained on the soundtracks of YouTube videos returned by searching on particular keywords were able to identify novel videos that had been tagged with the same keywords with surprising accuracy, even for tags (such as Parade) where the distinguishing acoustic characteristics are not completely clear. This opens the door to query-by-description search over large personal audio archives.

- Audio hashing/fingerprinting technologies, originally developed to match audio snippets against a large database of known music. can also be used to accelerate an exhaustive search for repeats in a large database. By using a fingerprinting scheme that relies on a small subset of the highest-energy components only, matching can occur despite variations in back-

ground noise and channel. Individual characteristic sounds can be very informative about a recorder's specific location.

- Our classifier-based music transcription remains competitive with more knowledge-based approaches, although all approaches have individual strengths and weaknesses. Extension to detecting multiple simultaneous piano notes works well, and can be improved simply by broadening and deepening the training corpus.

- Rather than attempting to identify precise notes, a "softer" representation of musical content, in the form of a vector of 12 "chroma weights" forms a more successful intermediate representation that preserves enough information about the original piece without hiding the uncertainties. This representation combined with beat-synchronous sampling is a very effective basis for matching different versions or performances of the same musical piece.

- Implementing our music signal analysis techniques within a more broadly targeted user interface, that allows arbitrary music recordings to be analyzed and investigated, shows great promise as a demonstration of applied science that can appeal to middle- and high-school children, possibly attracting them to further study in technical fields.

- Annotation of field recordings is a major obstacle to the analysis of animal communication; our experience in automatic environmental sound analysis is very relevant to these problems and can contribute major advances in the state of the art over current techniques used in these areas. In particular, we have developed a pitch tracker able to follow dolphin calls in underwater recordings at noise levels much higher than any previously-used technique.

2 Training and development

This year, the project has provided partial support for five graduate students: Ron Weiss, 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.

Support for the PI's involvement in the collaboration with the Computer Music Center, (Music Engineering Art Project) has further exposed a number of non-engineering graduate students to the techniques and procedures of engineering research.

Our work with high school student Jeff Bauer has helped encourage him towards an engineering career.

3 Outreach

This year saw presentations by the PI and students from LabROSA at Microsoft (Redmond), Ecole Normale Superiere (Paris), Danish Technical University, and Connecticut College, and conferences including the IEEE ICASSP (Toulouse), Neural Information Processing Systems (Vancouver), and a 2 hour invited tutorial at the Audio Engineering Society (Paris).

We also organized a successful workshop on Statistical and Perceptual Audition, SAPA-2006, which featured 12 presentations and attracted over 40 participants as a satellite to Interspeech 2006 in Pittsburgh.

The PI is currently involved in co-editing a special issue of IEEE Tr. Audio, Speech, and Language, on the topic of Music Information Retrieval. This issue should appear in 2008.

4 Contributions

4.1 Contributions to the principal discipline

We have contributed ideas and algorithms for the separation and identification of sound sources in real, complex environments based on their intrinsic properties as well as their spatial characteristics. We have also presented an approach for learning sound-to-word mappings through innovative exploitation of recently-available large data sources such as YouTube.

4.2 Contributions to sister disciplines

Our work in music signal analysis, particularly the automatic matching of songs with similar or matching melodic-harmonic structure, has great potential for impact and utility in musicology and music information science/archive management.

Our work in marine mammal sound analysis will, we hope, lay the foundation for large-scale objective analysis of dolphin communication calls, which can support a qualitative leap in the sophistication of our understanding of this behavior.

4.3 Contributions to human resources

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

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. Our work with music audio analysis, and in particular its realization in the MEAPsoft application, has potential impact on lay audiences by giving them an unprecedented perspective on the audio they enjoy by visualizing internal structure and similarity.

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, many updated and improved this year: `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 [5, 15, 10, 11, 8, 9, 12, 3, 13, 14, 7, 6, 4, 1].

References

[1] Adam Berenzweig. *Anchors and hubs in audio-based music similarity*. PhD thesis, Dept. of Elec. Eng., Columbia Univ., 2006.

[2] S. Downie, K. West, E. Pampalk, and P. Lamere. MIREX 2006 - Audio cover song. Online, 2006. URL http://www.music-ir.org/mirex2006/index.php/Audio_Cover_Song.

[3] D. P. W. Ellis and G. Poliner. Classification-based melody transcription. *Machine Learning Journal*, 2006. accepted for publication.

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

[5] D. P. W. Ellis and R. J. Weiss. Model-based monaural source separation using a vector-quantized phase-vocoder representation. In *Proc. ICASSP-06*, Toulouse, 2006. URL http://www.ee.columbia.edu/~dpwe/pubs/.

[6] Daniel P. W. Ellis. Beat tracking with dynamic programming. In *MIREX 2006 System Abstracts*, 2006. URL http://www.ee.columbia.edu/~dpwe/pubs/Ellis06-beattrack.pdf.

[7] Daniel P. W. Ellis. Identifying 'cover songs' with beat-synchronous chroma features. In *MIREX 2006 System Abstracts*, 2006. URL http://www.ee.columbia.edu/~dpwe/pubs/Ellis06-coversongs.pdf.

[8] Daniel P. W. Ellis and Keansub Lee. Accessing minimal-impact personal audio archives. *IEEE MultiMedia*, 13(4):30–38, Oct-Dec 2006. URL http://www.ee.columbia.edu/~dpwe/pubs/EllisL06-persaud.pdf.

[9] K. Lee and D. P. W. Ellis. Voice activity detection in personal audio recordings using autocorrelogram compensation. In *Proc. Interspeech*, pages 1970–1973, Pittsburgh, PA, Oct 2006. URL http://www.ee.columbia.edu/~dpwe/pubs/LeeE06-vad.pdf.

[10] M. I. Mandel and D. P. W. Ellis. A probability model for interaural phase difference. In *Proc. Workshop on Statistical and Perceptual Audition SAPA-06*,

pages 1–6, Pittsburgh, PA, Oct 2006. URL http://www.ee.columbia.edu/~dpwe/pubs/MandE06-probipd.pdf.

[11] M. I. Mandel, D. P. W. Ellis, and T. Jebara. An em algorithm for localizing multiple sound sources in reverberant environments. In *Proc. Neural Info. Proc. Sys.*, Vancouver, CA, Dec 2006. URL http://www.ee.columbia.edu/~dpwe/pubs/MandEJ06-EMloc.pdf.

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

[13] G. Poliner and D. P. W. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Applied Signal Processing - Special Issue on Music Signal Processing*, 2006. under review.

[14] G. E. Poliner, D. P. W. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE Tr. Audio, Speech, Lang. Proc.*, 2007. URL http://www.ee.columbia.edu/~dpwe/pubs/PolEFGSO07-meleval.pdf. accepted for publication.

[15] R. J. Weiss and D. P. W. Ellis. Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking. In *Proc. Workshop on Statistical and Perceptual Audition SAPA-06*, pages 31–36, Pittsburgh, PA, Oct 2006. URL http://www.ee.columbia.edu/~dpwe/pubs/WeissE06-rvm.pdf.

[16] Ron Weiss, Douglas Repetto, Mike Mandel, Dan Ellis, Victor Adan, Jeff Snyder, and Sam Pluta. MEAPsoft: Computers doing strange things with audio. Web site and software, Sep 2006. URL http://www.meapsoft.org/.