NSF-CAREER: The Listening Machine
Annual Report 2005

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

dpwe@ee.columbia.edu

Mar 21, 2006

1 Activities & Findings

1.1 Research Activities

Continuing our broadened theme of machine listening in many contexts, in 2005 we conducted research into automatic extraction of information in complex sound mixtures, in 'personal audio' environmental recordings, from music audio, and for the sounds of marine mammals recorded underwater.

1.1.1 Sound mixtures

2005 saw the graduation of Manuel Reyes, the Ph.D. student supported by this project from the start. Manuel's work on deformable spectrograms was rounded out by an application in which the model, which explains time-frequency distributions through a succession of transformations where possible, was applied in multiple subbands; where transformations cannot provide a good fit to the successive spectra, the signal is showing a more complex modification that could indicate the appearance of a new source. Thus, the signal can be divided into segments at such points, and individual sources pieced together by clustering segments that exhibit

common characteristics, including pitch and spectral distributions. This work was described in Manuel's thesis [10] and a submission to IEEE Trans. Speech & Audio [11].

Work on sound mixture organization and separation was continued by Ron Weiss who is looking at the use of signal models, in the form of vector quantizers, as a tool for identifying the contributions of individual sources amidst competing interference. In particular, we quantified the variation of model quality for speech as a function of codebook and training set size, looked at using perceptual spaces (e.g. mel frequency warping) for codebook design, and devised a scheme to store phase information (as phase advance) as part of the quantizer. This work was described in a submission to ICASSP-06 [4].

1.1.2 Personal audio

Our work in 'personal audio' concerns automatic techniques for identifying significant events in long recordings as might be collected by body-worn continuous recorders. Building on our earlier work to segment and cluster tens of hours of audio at a granularity of seconds, we recognized the particular importance of speech events in this data, and focused on the problem of identifying where speech occurs (which might be only a small portion of a day-long recording). This is essentially the Voice Activity Detection problem common in telephony, but here applied in far more challenging situations of high and variable background noise, and distant and possibly reverberated speech. Even when the speech is not intelligible, a listener can discern its presence, and we would like the automatic index to reflect this too. Based on the intuition that voice pitch is the most robust perceptual cue to speech, our approach focused on using a noise-robust pitch tracker to extract pitch despite high levels of background noise, and in distinguishing pitch due to voice from other kinds of periodicity (e.g. machines or music) based on dynamic characteristics. We described our work on personal audio archives in a submission to IEEE MultiMedia magazine's special issue on Continuous Archival and Recording of Personal Experiences [2].

We have also been looking at using spatial information from multiple microphones as a way of segregating information in environmental sound. Continuing our work with the Meeting Recorder, we hosted a visit by Kofi Boakye of Berkeley over the summer and worked with him on using cross-correlation between multiple microphone channels to identify and enhance the speech of particular participants. In our lab, Michael Mandel has begun working on using optimal probabilistic combination to infer spatial information from large amounts of noisy

observations, as is the case in reverberant environments. This work will be submitted for publication shortly.

1.1.3 Music audio

We have continued our successful work in extracting information from music audio. This year, we looked at the problem of extracting note events from music, firstly considering just the melody (i.e. most prominent notes) from conventional music recordings, then going on to look at identifying all the notes played in solo, polyphonic piano music. For both problems, we have used a radically simple classifier-based approach, which trains a Support Vector Machine on representative examples of audio frames for which the ground truth (actual melody or piano notes) has been prepared. We organized an evaluation of melody extraction algorithms as part of the 2005 International Conference on Music Information Retrieval (ISMIR-05) which attracted 10 submitted algorithms; we also prepared all the test materials and devised the evaluation metrics. Our work on transcription was described at ISMIR [8] as well as in submissions to Machine Learning Journal [3] and EURASIP [9].

We have also been working on a long-term project with a local high school student, Jeff Bauer, to investigate music listening habits among teenagers, one of the prime music consumer populations. We have devised a system to collect anonymous listening data from volunteers at his high school, with the intention of correlating listeners' reactions to specific songs, to see if we can describe different kinds of reactions to different music.

1.1.4 Marine mammals

As a still broader definition of machine listening, we have begun looking at the problem of organizing underwater sounds, specifically the vocalizations of marine mammals. Xanadu Halkias is a new student in the group with a passion for this topic, and the clear opportunities for applying speech and sound recognition techniques in this domain have encouraged us to pursue it. Xanadu made presentations at a special workshop on Marine Mammal Sound Analysis in Monaco in November 2005; she has been looking at the problem of clustering whale click sounds recorded by hydrophones in order to estimate the number of animals present [5], and more recently extracting and separating dolphin whistles from background noise and from each other (publication in preparation).

1.2 Educational activities

As in previous years, we have continued to develop two main courses: Digital Signal Processing (taught to seniors and first year masters students) and Speech and Audio Processing and Recognition (taught to more advanced graduate students). Both courses include a substantial project component which is devised by the student in consultation with me, and the projects continue to amaze me with their scope and quality (some examples are archived at http://www.ee.columbia.edu/~dpwe/teachingportfolio.html).

The Digital Signal Processing class is now also required for the imaging track of the Biomedical Engineering department in our school and we worked with that department to accommodate their students. In the Speech and Audio class, a new module on signal separation was developed and enhanced based on our most recent research. We also introduced material based on the personal audio application into the module on accessing large audio archives.

This year, our interdisciplinary collaboration with the Columbia Computer Music Center became much more concrete as I began teaching a weekly seminar, along with faculty from the CMC, in "Music Engineering Art Projects". With a core student body drawn equally from engineering and from the music and other departments, we have investigated the common ground in our techniques and approaches, as well as what we can gain from the complementarity of our disciplines. The current project is to develop an installation that will allow members of the Columbia community to bring their own music (e.g. from a portable music player), feed it into a booth, then have the machine re-render their music in a highly modified, but still hopefully recognizable and pleasing form. This process relies on using our music information extraction techniques to identify rhythm, tonality, and texture, and to extract and reorganize patterns in these quantities to regenerate music. While this work cannot be presented directly as engineering, it has certainly allowed us to explore a range of novel ideas – such as the feasibility of extracting a probabilistic description of tonality without specifically deciding on pitch – that will feed into more strictly-defined tasks like music retrieval. This ongoing collaboration has been extremely rewarding for all involved.

1.3 Findings

The findings from the research projects listed above include:

- Vector-quantizer type models present a simple and effective way to capture the intrinsic constraints represented by a specific signal source, and can be used as a basis for separating sounds in otherwise underdetermined situations.

- A noise-robust pitch tracker based on autocorrelation in subbands can successfully identify episodes containing speech, even when there are high levels of background noise and/or reverberation. Tracking the pitch further offers opportunities to enhance the speech via harmonic filtering.

- Identifying the spatial location of sources in reverberant environments is feasible as long as the weak information is advantageously combined e.g. through applications of probability theory. Advanced techniques such as particle filtering can make this inference process relatively tractable.

- The classifier framework for music note transcription is competitive with traditional model-based approaches (our system was a few percent worse than the best in our evaluation). Moreover, the same approach trivially extends to extracting full polyphony for pianos, resulting in transcriptions that far outperform previously published approaches in a direct comparison.

- Clustering techniques such as spectral clustering can work well for automatically organizing the complex sounds from groups of marine mammals recorded underwater. Probabilistic approaches also help in the disambiguation and separation of overlapping whistle calls e.g. from multiple dolphins.

2 Training and development

This project provided support for graduate student Manuel Reyes through to his graduation in May, and thereafter to Ron Weiss who has taken over work on model-based signal separation. Ron was an undergraduate in our department and I am very pleased to have him join my group.

Partial support has again been provided to Keansub Lee, working on the personal audio project. This year saw Keansub's transition from working on projects as directed to devising and developing his own ideas for the next stages in this project.

The Music Engineering Art Project has involved 8-10 graduate students from various different departments throughout the year. Because of the interdisciplinary nature of the group, everyone has learned a great deal, both in terms of specific technical details (of signal processing or music theory), and in a broader sense of relating to and understanding different approaches to academic research.

Our work with high school student Jeff Bauer and undergraduate student Suman Ravuri has helped induct and guide these younger people towards the academic path.

3 Outreach

This year saw presentations by the PI and students from LabROSA at Google (New York), Univ. Oldenburg (Germany), Rutgers Univ., Johns Hopkins Univ., and conferences including the International Conference on Music Information Retrieval (ISMIR), IEEE ICASSP (Philadelphia), and the opening keynote at the IEEE Workshop on Automatic Speech Recognition and Understanding (Puerto Rico).

We also worked on editing a special issue of IEEE Tr. Speech and Audio, following on the themes of the 2004 Workshop on Statistical and Perceptual Audio Processing that we organized. Papers were submitted in January 2005; in conjunction with my co-editors, we supervised the review of 26 papers and ended up accepting 12 to appear in a 2006 edition of the journal. We are planning a follow-on workshop in 2006.

4 Contributions

4.1 Contributions to the principal discipline

We have contributed novel algorithms for signal separation and organization both through our academic publications, and through software releases made on our web site. In particular, this year we described novel approaches to spectral quantization and reconstruction, to detecting speech in high-noise environments, and to identifying the spatial location of reverberated sources.

4.2 Contributions to sister disciplines

We have made significant contributions to music information processing, both through our development of the novel classifier-based transcription technique, and through our organization of the international evaluation of melody transcription systems. We have contributed to marine science by developing techniques for automatic extraction of biologically-significant information from underwater sounds, devised in response to the needs of marine researchers we have contacted through relevant meetings and direct approaches.

4.3 Contributions to human resources

The project has provided direct (partial) support for three graduate students, and has enabled the PI to participate in research supervision of two more graduate students, one undergraduate, and one high school student, all of whom worked primarily on their own individual research projects.

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

This year we developed the idea of applying signal separation techniques to problems beyond the narrow realm of artificial signal mixtures and objective metrics,

to look at real-world problems and delivering benefits to real listeners. We submitted a proposal to the NSF, in collaboration with other engineers and psychologists, to develop signal separation specifically aimed to improve speech intelligibility. This task is difficult both to achieve and to measure, but strikes us as an important step in closing the loop between abstract research and practical benefits. This project will begin in 2006.

Our music information extraction work is similarly inspired by the desire to solve real problems – helping listeners connect with the music they want to listen to. Music technology is a rapidly moving field, and this year we have been in contact with several commercial organizations looking to develop mass-market products incorporating some of these ideas. Fruits of these exchanges will likely appear in 2006.

5 Resources made available

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

- 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/
- Class notes and session summaries from the ongoing seminar in Music Engineering Art Project: http://works.music.columbia.edu/MEAP/
- 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/

6 Publications

See references [9, 3, 7, 8, 1, 2, 10, 11, 6].

References

[1] K. Dobson, B. Whitman, and D. P. W. Ellis. Learning auditory models of machine voices. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics WASPAA*, Mohonk NY, Oct 2005.

[2] D. P. W. Ellis and K. Lee. Collecting and using minimal-impact personal audio archives. *IEEE MultiMedia Magazine, special issue on Recording Personal Experiences*, 2005. under review.

[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 R. J. Weiss. Model-based monaural source separation using a vector-quantized phase-vocoder representation. In *Proc. ICASSP-06*, Toulouse, 2006.

[5] X. Halkias and D. P. W. Ellis. Estimating the number of marine mammals using recordings of clicks from one microphone. In *Proc. ICASSP-06*, Toulouse, 2006.

[6] N. Lesser and D. P. W. Ellis. Clap detection and discrimination for rhythm therapy. In *Proc. IEEE ICASSP-05*, Philadelphia PA, March 2005.

[7] M. Mandel and D. P. W. Ellis. Song-level features and support vector machines for music classification. In *Proc. International Conference on Music Information Retrieval ISMIR*, London, Sep 2005.

[8] G. Poliner and D. P. W. Ellis. A classification approach to melody transcription. In *Proc. International Conference on Music Information Retrieval ISMIR*, London, Sep 2005.

[9] 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.

[10] M. J. Reyes-Gomez. *Statistical Graphical Models for Scene Analysis, Source Separation and Other Audio Applications*. PhD thesis, Dept. of Elec. Eng., Columbia Univ., 2005.

[11] M. J. Reyes-Gomez, N. Jojic, and D. P. Ellis. The deformable spectrogram model for sound dynamics. *IEEE Tr. Speech and Audio Proc.*, 2005. under review.