NSF-CAREER: The Listening Machine
Annual Report 2004

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

dpwe@ee.columbia.edu

Jan 02, 2005

1 Activities & Findings

1.1 Research and Education Activities

This year, we expanded our investigation of sound analysis to look at a range of different sound 'scenes', including overlapping conversations, ambient everyday sound, and music. In each case, the goal is to abstract useful information similar to that which a human listener would perceive, and in particular to deal successfully with the issues raised by multiple, overlapping sound sources.

Our most focused effort was the continued development of the novel model for sound sources we proposed last year, based on treating each spectral instant as a simple deformation of its immediate predecessor (or, in general, its neighbors). This model decomposes smoothly-varying segments of sound into a single spectral profile, and a set of locally-smooth transformation functions, describing how the spectral detail is derived from its predecessors. This year, we extended this model to a two-layer version with separate transformations applied to fine spectral structure (e.g. harmonics, to account for changes in pitch) and broader spectral structure (e.g. the formants of voice, which in general will move independently of the harmonics). The key to this model is the way that the parame-

ters for the generative model are inferred from a particular signal, which involves an approximation to exact Bayesian inference based on belief propagation. This model has numerous applications which we have only just begun to investigate: it can be used to interpolate across missing segments of signal (while preserving the independent dynamics of pitch and formants); the transformation parameters themselves constitute a promising and noise-robust feature for speech and perhaps speaker recognition; and the points at which signals cannot be adequately represented as transformations (in which case the frame is fit to a static dictionary) represent likely boundaries between different sources in mixed, overlapping signals. This work, which was performed in collaboration with Microsoft Research, is described in [6, 7].

This year we initiated the collection and processing of 'personal audio' – sound collected by small body-worn devices that record continuously. As a result of the explosion in personal audio technology, small light-weight recorders that can capture 12 h or more of audio are readily available, but navigating and indexing such recordings remains an enormous challenge. We collected a small database consisting of 60 h of data that was then manually labeled to mark the major 'episodes' i.e. changes in activity or location, corresponding to the kind of detail that might be entered in an on-line calendar. Our data includes 139 segments (average duration: 26 mins), which have been sorted into 16 classes such as "library", "class", "restaurant" etc. As a first investigation into a possible useful application for these kinds of recordings, we attempt to recover these episodic labels, which involves two tasks: finding the segmentation points between internally-consistent regions, then clustering each of the resulting segments into classes that appear to contain repeats of the same activity or location. Using ideas from speaker segmentation and clustering, we found boundaries based on a Bayesian Information Criterion, and clustered the segments with the Spectral Clustering algorithm, applied to a matrix of Kullback-Liebler distances between feature vector distributions in each segment. We experimented with different feature vectors (21-band auditory log-spectra were best) and granularity of the analysis frames, which involves a trade-off between using segments long enough to get consistent statistics for episode-level events, without blurring over changes in events. We found segments in the range 3-30 s to be most effective. Our best results achieved around 80% correct labeling of frames under the optimal assignment of anonymous clusters to ground truth labels. Our work was described in [2, 3].

A very specific type of sound event analysis was pursued in a project to develop 'rhythm therapy' interface. Although we did not have a formal connection

to the supplier, a student became interested in a therapy for attention and focus improvement in which children clap in time to a visual or headphones-presented stimulus. Clinical results have shown that improving time accuracy down to 10s of ms is correlated with improvements in concentration and other behaviors. Current feedback systems rely on a special-purpose mechanical sensor held in the hand, but in theory the sound of the clap should provide sufficient information to be detected with the microphone typically already built-in to a standard laptop computer. At the same time, if multiple students are simultaneously engaged in the therapy in the same room, there is potential confusion between the claps of adjacent users. The task, then, was to build a system to discriminate between 'near-field' claps (e.g. from within 1 m of the microphone) and more distant, or off-axis clap sounds. By devising some signal measures that correlate very well with the standard psychoacoustic cue to source distance, the direct-to-reverberant ratio, we were able to get near perfect discrimination. As part of the project, we collected a database of several thousand claps for different source and microphone locations in different rooms. Our work is described in [5], although we have continued work since them in response to the very promising results we saw.

Our parallel work on considering music audio has also continued. By analogy to the work above, we are interested in identifying episodes and events in recorded music. We are investigating several directions: One, based on [4], first finds onsets in the audio, then tries to cluster those onsets into similar sounds by successively combining similar onsets into a template, then searching through the audio to find instances of that template. We found this approach to work well for percussion sounds with a few strong spectral peaks (resonant modes) but otherwise problematic. We are now investigating a more general 'hidden event' approach to this problem, where both onset times and time-frequency structure are jointly estimated to optimize a description of the audio. The second (but related) approach is based on a fingerprinting algorithm used to identify small snippets of existing recordings, as might be heard on the radio [8]. It finds a large number of 'hash elements', derived from pairs of local peaks in a time-frequency surface and their relative timing. Any given piece of music will generate a large number of hashes, many of which will be robust to filtering (reverberation), added noise, etc. While any single hash is not very discriminant, finding a whole sequence of such elements in the same relative timing as in a reference piece rapidly becomes a very accurate indicator of a match. Our interest is to apply approach of using a greatly reduced signature representation not to identify repeats of whole songs, but to identify repeating sounds within audio (e.g. repeating notes on the same instrument within a particular piece of music). Surprisingly, the original fingerprint

features can vary quite significantly even between musical excerpts that sound identical, such as the repeated choruses of a pop song (which perhaps explains why the approach is so useful in fingerprinting whole songs), but we are looking at how to reweight the components of the hash to getter a better correspondance between perceptual resemblance and hash hits.

The interest in percussion events is based on our earlier work on modeling the drum patterns in popular music through basis functions, which has thus far been based on symbolic descriptions (scores and captured performances) rather than acoustic input. Given the set of timings corresponding to the drum tracks, we normalize for tempo and downbeat time to produce a 'stack' of aligned patterns which can then be condensed to a set of basis functions by one of several standard mechanisms. We investigated principal component analysis, independent component analysis, linear discriminant analysis, and non-negative matrix factorization. Each of these gave interesting variants on a set of drum pattern 'parts' that can be combined to give an infinite range of musically-meaningful drum patterns; analyzing a particular pattern into these functions might be useful for genre or style classification [1].

Also on the music front, we have started looking at the long-standing problem of recovering a score from music audio – the transcription problem. This work began when an evaluation was organized as part of last year's International Conference on Music Information Retrieval (ISMIR) to compare different algorithms for recovering the 'dominant' melody track (such as the singer's notes) in a recording. Our idea was to cast this as a pure classification problem – identify the correct melody-note label given the spectral feature vector – rather than trying to build-in prior knowledge about the realization of pitches in spectra as in previous models. Our submission performed comparably to the 3 other algorithms (from different countries) submitted to the evaluation – though not best, our system was considerably simpler and did not involve any post-processing to remove outliers. A paper co-authored by all participants is under preparation. This project also involved a high-school student working at the lab as a summer intern, who tackled the problem of identifying the melody line in a machine-readable score (a MIDI file) to extract training data for our classifier; he ended up submitting this project to the Intel Science and Engineering Fair; he won the local region and is currently on his way to Arizona for the national finals.

1.1.1 Educational activities

Our research activities continue to feed into curricular materials for the senior/masters level Digital Signal Processing course http://www.ee.columbia.edu/~dpwe/e4810/, and the advanced graduate course on Speech and Audio Processing and Recognition http://www.ee.columbia.edu/~dpwe/e6820/. The full online course materials (slide packs, sound examples, matlab demos) have been progressively extended and refined over last year. The suite of online Matlab examples of instructive audio processing examples continues to be developed; additions this year included:

- Linear and log-frequency spectrograms, including routines to convert between the two. Log-frequency time-frequency representations are closer to the analysis performed by the ear, and are a natural representation for music, whose harmonies are based on ratios. http://www.ee.columbia.edu/~dpwe/resources/matlab/sgram/

- Improvements to the example of warping spectra by manipulating their all-pole (LPC) representation, including online examples. http://www.ee.columbia.edu/~dpwe/resources/matlab/polewarp/

- A utility to read soundfiles compressed using the ubiquitous MP3 format directly into Matlab. Astonishingly, no existing tool for this was available (as far as we could tell). http://www.ee.columbia.edu/~dpwe/resources/matlab/mp3read.html

We have established a new collaboration with the Computer Music Center at Columbia, where researchers are working on many of the same audio technology issues that interest us, albeit with different goals and criteria. We are currently planning a new projects course to be run as a collaboration between our two schools where music and engineering students will work together on projects of mutual interest. We hope for unexpected benefits from exposing our engineers' mindsets to the more open-ended approaches common in aesthetic endeavors. We also recognize a growing demand from students not content to limit themselves within traditional arts or engineering categories, students for whom technical sophistication and computational tools are second nature, regardless of their particular orientations. http://works.music.columbia.edu/MEAP/.

1.2 Findings

The findings from the research projects listed above include:

- The deformable spectrogram model has proved an exciting and novel way to describe audio signals. In particular, the ability to identify segmentation points (e.g. where dominant sources change) and to interpolate across missing regions holds great promise for signal separation and restoration (and will form the basis for a separate forthcoming proposal).

- We have shown that a calendar-like representation can indeed be recovered with good accuracy from easily-collected ubiquitous personal audio recordings.

- Clapping in a normal room can be classified according to its range with very high accuracy, by using a simple measure to capture the direct-to-reverberant ratio; it would be very practical to devise audio-only clap detectors, and these could be used by several people in the same room, provided a minimal spacing between users was achieved. Further investigation will reveal how well this technique applies to less impulsive sounds.

- Expressing musical drum patterns in terms of basis functions leads to meaningful and interesting insights into the structure and correlation of different parts of the patterns, and could have applications in both music synthesis and stylistic categorization.

- Musical melody transcription can be achieved within a standard classifier framework making no assumptions about the physical realizations of a particular pitch. The training data for this classifier can be derived from widely-available MIDI renditions, which provide both a machine-readable transcription and the associated audio waveform.

2 Training and development

This project provided full academic year support for one graduate student, Manuel Reyes, who developed the deformable spectrogram work described above. Manuel has continued to make very well-received conference presentations and is planning several journal publications this year. He will likely defend his Ph.D. in the summer.

Partial support has also been provided to Keansub Lee, who is working on the personal audio project. Keansub has developed enormously through this project, and is developing very good academic habits, including deep research into existing publications and systematic investigations into algorithm variations.

The music work continues to be an excellent way to get students involved in and excited about research. For the second year, we participated in the New York Academy of Sciences summer internship program, leading to the Intel Science and Engineering Fair prize for highschool junior Ben Chang as mentioned above. Ben developed excellent independent research and presentation skills. Three other graduate students, Michael Mandel, Graham Poliner, and Ron Weiss, while not supported directly by this project, were involved in the research on music signal analysis.

3 Outreach

This year saw presentations by the PI and students from LabROSA at Microsoft (Redmond), Queen Mary (University of London), ACM Multimedia, International Conference on Music Information Retrieval (ISMIR), Acoustical Society of America, and IEEE ICASSP.

The PI was involved in organizing several workshops this year. First was the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-2004 http://www.sapa2004.org/, held as satellite to ICSLP-2004 in Korea in October 2004. 22 papers were presented at the workshop covering a wide but rich range of topics in innovative audio processing; we are currently editing a special issue of IEEE Transactions on Speech and Audio Processing related to this meeting.

Second was a follow-on to 2003's NSF-sponsored workshop on speech separation. This year, additional funding was obtained from AFOSR for the workshop on 'Speech Separation and Comprehension in Complex Acoustic Environments' http://labrosa.ee.columbia.edu/Montreal2004/, held over three and a half days with 40 invited participants. This was a very well received forum for engineers, computer scientists, psychologists, and physiologist to share ideas over their common interest in the separation of speech sounds from acoustic interference. Particularly successful was the inclusion of 10 students/postdocs who made poster presentations of their ongoing work.

The third workshop concerned Music Information Processing Systems, held as one of the workshops at 2004's Neural Information Processing Systems meeting http://www.iro.umontreal.ca/~eckdoug/mips/. This one-day workshop involved 8 presentations and several extended discussions on the emerging area of applying machine learning and other statistical techniques to music analysis and processing.

4 Contributions

4.1 Contributions to the principle discipline

Through published articles and software resources made available on the web, we have contributed a number of new algorithms for extracting information from acoustic signals, including the deformable spectrograms, the use of long-window features, BIC segmentation, and spectral clustering on long-duration recordings, and trained classifiers for music signal transcription.

4.2 Contributions to sister disciplines

Through our new collaboration with the Music department, and through our participation in forums such as the International Conference on Music Information Retrieval, we have made several technical contributions to related musical fields.

4.3 Contributions to human resources

The project has provided direct (partial) support for two graduate students. It has also enabled the PI to participate in research supervision of three more graduate students, one MS student, and one highschool junior, 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 online educational materials. We continue to receive positive feedback from academics all over the world who have found these materials useful, and have asked for (and obtained!) permission to adapt them to their own uses.

4.5 Contributions to wider aspects of public welfare

We are very interested in developing techniques with applicability in real, practical applications with wide utility. The deformable spectrogram model holds great promise for signal separation and restoration, and we hope to develop this into the foundation of an algorithm that could enable future “super hearing aids”, useful even to normal-hearing listeners experiencing extremely noisy conditions. The personal audio diary project is again motivated by the wish to provide a way to

allow people to gain value from long-duration personal recordings that they can already make, if only there was some value in doing so. The music work is marshalled behind an overarching goal of automatic music similarity inference, which would have a very profound influence on helping to connect listeners to the music they will enjoy, regardless of how popular it is in the mainstream.

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/
- Focused collection of Sound Examples for use in student projects: http://www.ee.columbia.edu/~dpwe/sounds/
- Matlab examples of common audio processing algorithms, including log-frequency spectrograms, and reading MP3 files, new this year: http://www.ee.columbia.edu/~dpwe/resources/matlab/

6 Publications

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

References

[1] D. P. W. Ellis and J. Arroyo. Eigenrhythms: Drum pattern basis sets for classification and generation. In *Proc. Int. Symp. on Music Info. Retr. ISMIR-04*, Barcelona, October 2004.

[2] D. P. W. Ellis and K. Lee. Features for segmenting and classifying long-duration recordings of "personal" audio. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04*, Jeju, Korea, October 2004. http://www.ee.columbia.edu/~dpwe/pubs/sapa04-persaud.pdf.

[3] D. P. W. Ellis and K. Lee. Minimal-impact audio-based personal archives. In *Proceedings of the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, New York, NY, October 2004.

[4] M. G. Kazuyoshi Yoshii and H. G. Okuno. Drum sound identification for polyphonic music using template adaptation and matching methods. In *Proc. Workshop on Statistical and Perceptual Audio Proc. SAPA-04*, Jeju, Korea, October 2004.

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

[6] M. Reyes-Gomez, N. Jojic, and D. P. W. Ellis. Towards single-channel unsupervised source separation of speech mixtures: The layered harmonics/formants separation-tracking model. In *Proc. Workshop on Statistical and Perceptual Audio Proc. SAPA-04*, Jeju, Korea, October 2004.

[7] M. Reyes-Gomez, N. Jojic, and D. P. W. Ellis. Deformable spectrograms. In *Proc. AI and Statistics*, Barbados, 2005.

[8] A. Wang. An industrial strength audio search algorithm. In *Proc. Int. Conf. on Music Info. Retrieval ISMIR-03*, 2003.