The Listening Machine: 1st Annual Report

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

dpwe@ee.columbia.edu

March 15, 2004

1 Activities & Findings

1.1 Research and Education Activities

In this first year of the project, our work was focused on the problem of identifying and separating specific sound sources in mixtures. The core of our approach is to use prior knowledge about the sounds in the world, encapsulated in some kind of model, to provide the constraints needed to solve the blind separation problem which is otherwise ill-posed.

We have looked at using this approach in a reverberant multi-microphone case. In collaboration with Bhiksha Raj of MERL in Cambridge, we looked at setting the parameters of a filter-and-sum beamformer by doing gradient descent on the match between the separated signals and the constrained speech approximation resulting from the model means corresponding to the states of the best-match path found by a speech recognizer [9]. Beam-former parameters and speech recognizer state path parameters can be alternately re-estimated; we found this process to converge successfully after a few cycles. When just a single voice is present, this process amounts to blind estimation of a dereverberation filter. But we were more interested in the problem of multiple overlapping voices, which requires two initial speech recognizer state paths. This required a factorial-HMM model for the mixed speech, where both states as well as the parameters of an unknown linear state-mean mixing matrix are inferred by variational EM.

In many cases, multi-microphone signals are not available, so we have also looked at the single-channel separation case. Here, instead of using a model for speech borrowed from a speech recognizer, we trained a separate hidden Markov model of speech, which can then be optimized for the task of estimating the speech spectrum (rather than recovering phonetic states). Starting from the model of [10], we found that a single-chain state-based model required thousands of states for accurate representation of the spectral variation in a high-resolution spectrum (which can then be used to separate the speech as a gain mask applied to the short-time Fourier transform). Since the state transition matrix is $O(N^2)$ in the number of states, model learning and state path inference rapidly become intractable for these large models. Instead, we have been looking at ways to factor these models into several loosely-coupled factors which can be separately modeled with smaller numbers of states. In collaboration with Nebojsa Jojic of Microsoft Research, our first realization of this idea breaks the signal down into subbands consisting of between 19 and 64 FFT bins, then trains separate but coupled hidden Markov models for each subband [7]. The state evolution in each band depends on the features within that band, but also on the states in the two adjacent bands; variational inference is performed on whole bands sequentially, taking the estimated state posteriors (the ‘variational parameters’) from the most recent iteration of the adjacent channels as fixed. The models converge in fewer than 20 passes over the full spectrum of bands.

Our most recent approach, again in collaboration with Microsoft, attempts to improve the accuracy of signal models by incorporating earlier observations of the actual signal through the operation of local transformation operators, for instance modeling each FFT bin as resulting from one of several possible linear functions applied to a local region of bins in the previous frame, where the functions are currently preset to allow for translations up and down in frequency, but eventually we aim to learn the functions from the data. Transformations are estimated for an entire signal, subject to local consistency constraints applied via belief propagation [8]. This model can be used directly to interpolate and restore missing or obscured regions of an audio signal, but more generally provides a richer basis for detailed models of audio signals.

Other work this year has looked at the application of signal models for recognition of partially-observed sources. Building on previous work applied in the speech domain [1] and our continuing research into detecting alarm sounds [3], we have looked at different domains and models for describing alarm sounds that facilitate their detection regardless of background noise characteristics. Our approach has been to attempt to learn spectral models based only on the peak frequencies of noisy sounds, using synthetic alarm-noise mixture in which we know precisely where the alarms occur, so we can perform supervised model learning. However, we have found that the 'auditory' spectral representation we have been using (based on the Bark-scaled filterbank of [4]) does not offer adequate spectral resolution to capture the characteristic details of alarm sounds, which often include sparse, steady harmonics. At the same time, moving to a finer spectral representation would defeat our ability to generalize between different alarm examples. We are continuing this work by looking at features such as within-band spectral entropy which can distinguish between concentrated and diffuse spectra without actually having to sample the spectrum more densely.

The final research thread concerns music analysis, since we are also interested in applying our sound organization techniques in that domain. Although this work is at an early stage, one of the first issues is obtaining detailed information concerning the events within music, and we have previously investigated an approach of aligning music recordings to simplified 'replicas' created by amateur musicians in the MIDI format [11]. On the current project, we worked with a summer intern from a local high school to build a database of hand-marked segments in real music which can be used to train and test future automatic analysis systems. Although not directly applicable to analysis of the signal content in terms of particular sound sources, we have also continued the development and evaluation of our database of ground truth in music similarity, which, spanning 400 popular artists and almost 9,000 individual tracks, is the largest and best-studied database of its kind.

1.1.1 Educational activities

On the education side, the past year included refinements and additions to the core introductory Digital Signal Processing class and the graduate-level Speech and Audio Processing and Recognition class, and the development of a new mini-course on Music Content Analysis by Machine Learning. In addition, the PI has established a new seminar class on Machine Learning for Signal Processing, http://labrosa.ee.columbia.edu/wiki/rg?SeminarHome .

For the introductory Digital Signal Processing course, the entire course materials, including over 400 pages of powerpoint slides, have been made available on the course web site, http://www.ee.columbia.edu/~dpwe/e4810/ . The 450 slides for the Speech and Audio Processing and Recognition class are similarly available from the class web site, http://www.ee.columbia.edu/~dpwe/e6820/ . The PI has received notes of thanks, and been happy to grant permission to use these slides from a number of colleagues all over the world.

Additional materials of broader relevance created specifically for class demonstrations and to support the individual student projects in both classes include:

- A tutorial/implementation of sound spectral warping based on bilinear transformation of linear prediction models. http://www.ee.columbia.edu/~dpwe/resources/matlab/polewarp/

- A comprehensive, modular Matlab reimplementation of the popular Rasta-PLP speech front-end feature calculation algorithm, which was taken up in several projects.

- A diverse, categorized database of sound examples including voice, musical instruments, natural sounds and field recordings to provide source material and inspirations for student projects. http://www.ee.columbia.edu/~dpwe/sounds/

A new short course on Machine Learning for Music Analysis was created in response to an invitation for a one-week visit Pompeu Fabreu University in Barcelona, and further refined to include an extensive self-paced practical for a summer school at Johns Hopkins University. The course covers the basics of pattern recognition as applied to audio signals, illustrated with the task of locating stretches of singing voice in popular music recordings. The slides and practical are available online: http://www.ee.columbia.edu/~dpwe/muscontent/ .

1.2 Findings

The findings from the research projects listed above include:

- In the multi-mic reverberant speech separation case, we found that (a) gradient descent to match the processed outputs to targets estimated by a speech recognizer was a successful separation algorithm; and (b) approximate inference by variational methods on a factorial hidden Markov model assuming a simple linear combination rule was able to find suitable gradient-descent targets for recordings of two simultaneous voices even when the mic locations and reverberation conditions were unknown.

- Models of speech signals consisting of a collection of separate Markov chains covering separate frequency subbands, and coupled to their immediate neighbors through transition probabilities were able to match speech signals accurately and efficiently; these models achieved satisfactory signal-to-noise ratio improvements when separating overlapping voices via time-frequency masking.

- It is feasible to describe time-frequency energy distributions for sound in terms of transformations, where each frame is partially explained as a composition of frequency-local translations. After learning the appropriate constraints for this model from real sound, it can be used restore deleted portions of the sound which, while unlikely to be a close match to the deleted portion, never the less sound very reasonable in the context of a resynthesis.

- In trying to automatically detect alarm sounds over a wide range of background sounds, it is rarely adequate to model just then spectral magnitude on an auditory-scaled basis (such as MFCCs or their spectral equivalent). We believe finer frequency resolution is required, and are investigating additional features to provide some of this information.

- The main section breaks in a given piece of popular music can be manually marked with relative ease (to provide a useful database of ground-truth for real-world acoustic recordings).

2 Training and development

This project provided full academic year support for one graduate student, Manuel Reyes, who developed the main signal separation and modeling work described above. With three significant publications accepted in the past 12 months, Manuel is rapidly becoming a well-respected, full participant in the scientific community.

Summer support was also provided for a second student, Uday Arya, who has been working on the alarms project. Uday has gained his first experience of independent research in this project, and has made good progress in the areas of designing and conducting systematic empirical investigations.

Also involved over the summer (through an internship program sponsored by the New York Academy of Sciences) was Angel Umpierre, a high school senior from a local public school. Angel gained insight into the practice of science and research, and made a valuable contribution in terms of his hand-marked ground-truth data for music analysis and segmentation.

3 Outreach

In addition to the external teaching at UPF Barcelona and Johns Hopkins mentioned above, the PI and the students involved have given talks presenting results from this work at MIT, Stanford, the University of Pennsylvania, the City University of New York, Google, Microsoft Research, the Acoustical Society meetings in Nashville and Austin, IEEE International Conference on Multimedia ICME-03 in Baltimore, and the IEEE Workshop on Applications of Signal Processing to Audio in New Paltz.

The PI was also a co-organizer (with Pierre Divenyi and DeLiang Wang) of an NSF-sponsored meeting, "Perspectives on Speech Separation" held in Montreal at the end of October 2003 (IIS-0345301). This invitation-only interdisciplinary workshop brought together behavioral, computational, and neuroscientists from the U.S. and abroad for an in-depth discussion of the current state of research on speech separation by humans, machines, and the nervous system; details including presentations are available online at http://www.ebire.org/speechseparation/ . Proceedings of the workshop will appear in a book to appear this year [2].

The PI is also a co-organizer (with Bhiksha Raj and Paris Smaragdis) of the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-2004 http://www.sapa2004.org/, to be held as satellite to ICSLP-2004 in Korea in October 2004. This workshop is intended to bring together researchers from a broad range of disciplines including blind source separation, auditory modeling, and speech recognition, to discuss different approaches and solutions to acoustic signal separation.

4 Contributions

4.1 Contributions to the principle discipline

The project has already developed a range of novel models and techniques for describing, recognizing, separating and reconstruction natural and man-made sounds in complex, real-world environments. Specifically, we have proposed and investigated several novel signal processing and machine learning structures based on multi- and single-channel systems using a variety of recently-developed learning techniques including factorial and coupled hidden Markov models, variational approximations, and belief propagation.

4.2 Contributions to sister disciplines

Complex signals and cluttered scenes occur in many analogous disciplines including vision, underwater acoustics, radar, etc. Our techniques and models may find application in these fields.

The project has had a secondary impact on research into music signal analysis being carried out at the lab, including labeling contributions of the summer intern.

4.3 Contributions to human resources

The project has directly supported the ongoing professional development of one Ph.D. candidate, and, through summer internships, on Masters candidate and one high-school student who were given their first opportunities to perform individual work on a real research project.

4.4 Contributions to educational infrastructure

The project has supported the development of a range of educational materials including class notes for three classes and one seminar (all freely available over the web), and various supporting data and tutorial implementations created for class projects and forging the link between classroom and research.

4.5 Contributions to wider aspects of public welfare

Our ultimate goals with this project include the development of intelligent automatic media processing, able to help information consumers browse and search the currently opaque archives of audio and audio-visual materials. Practical machine listening techniques will also be required by future autonomous machines that can navigate in the real world making used of the same environmental cues available to people. Finally, our spin-off work in music similarity has the potential to help all music lovers more easily find a wider range of music to their liking.

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/

- Reading list and session summaries from the ongoing seminar in Machine Learning for Signal Processing: http://labrosa.ee.columbia.edu/wiki/rg?SeminarHome

- 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 sections on voice warping via bilinear mapping, dynamic time warping, and Rasta-PLP, all new this year: http://www.ee.columbia.edu/~dpwe/resources/matlab/

- Hand-marked transcriptions of major boundaries in real pop music examples: http://www.ee.columbia.edu/~dpwe/sounds/music/

- Database definition and range of subjective similarity ground-truth datasets for 400 popular music artists: http://www.ee.columbia.edu/~dpwe/research/musicsim/

6 Publications

See references [5, 6, 7, 8, 9].

References

[1] J. Barker, M. Cooke, and D. P. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 2004. Submitted.

[2] P. L. Divenyi, editor. *Perspectives on Speech Separation*. Kluwer Academic Publishers, New York. to appear summer 2004.

[3] D. Ellis. Detecting alarm sounds. In *Proc. Workshop on Consistent and Reliable Acoustic Cues CRAC-2000*, Aalborg, 2001.

[4] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *JASA*, 87(4):1738–1752, apr 1990.

[5] M. Reyes-Gomez and D. Ellis. Selection, parameter estimation, and discriminative training of hidden markov models for general audio modeling. In *Proc. IEEE International Conference on Multimedia and Expo ICME-03*, Baltimore, July 2003.

[6] M. Reyes-Gomez, B. Raj, and D. Ellis. Multi-channel source separation by factorial hmms. In *Proc. IEEE ICASSP-03*, Hong Kong, 2003.

[7] M. J. Reyes-Gomez, D. P. Ellis, and N. Jojic. Multiband audio modeling for single channel acoustic source separation. In *Proc. ICASSP-04*, Montreal, 2004.

[8] M. J. Reyes-Gomez, N. Jojic, and D. P. Ellis. Interpolating audio signals with transformational models. Submittedd to 2004 Snowbird Learning Workshop.

[9] M. J. Reyes-Gomez, B. Raj, and D. P. Ellis. Multi-channel source separation by beamforming trained with factorial HMMs. In *Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk NY, 2003.

[10] S. Roweis. One-microphone source separation. In *Advances in NIPS*, pages 609–616. MIT Press, Cambridge MA, 2000.

[11] R. J. Turetsky and D. P. Ellis. Ground-truth transcriptions of real music from force-aligned midi syntheses. In *Proc. Int. Conf. on Music Info. Retrieval ISMIR-03*, Baltimore MD, 2003.