NSF-CAREER: The Listening Machine
IIS-0238301 2003–2008
Final Report

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

dpwe@ee.columbia.edu

June 4th, 2009

1 Activities & Findings

1.1 Research Activities

This six-year project started with the idea of applying sound recognition and separation techniques that had originated in speech recognition to a broader domain of environmental sound mixtures. As it proceeded, the work diversified into several distinct areas, reflecting the different directions of the graduate students primarily supported by the project: Manuel Reyes and Keansub Lee worked on environmental sound separation and classification; Graham Poliner and Michael Mandel worked on music audio transcription and classification, and Xanadu Halkias applied many of the same ideas and techniques to the analysis of underwater recordings of dolphin sounds.

1.1.1 Sound Separation

A key problem in analysis and classification of sound mixtures is that the sounds emanating from individual sources cannot, in general, be directly observed or measured, only the combined superposition of all simultaneous sources. To address this, we looked at a number of techniques for separating or describing individual sources in mixtures, particularly in single-channel (underdetermined conditions), although our earliest work dealt with a reverberant multi-microphone scenario in which the transcription of two voices was achieved by searching for beamforming filter parameters that maximized the total likelihood of a pair of speech state estimates from a factorial hidden Markov model (FHMM). Using the likelihood of a speech recognizer as the objective function differentiates this work from the more common approach using signal energy or higher-order moments, and allows the system not only to separate the voices, but also to make them adapt to the recognizer model e.g. by suppressing reverberation [32, 33].

Our subsequent work focused on the single-channel case, where several voices are superimposed in a single microphone recording. The key to solving this underconstrained problem is to exploit the known structure of the speech signal by applying models trained to recognize the particular structures of speech. Since this always involves a compromise between accuracy and model size, we investigated breaking it down by having separate models for multiple frequency subbands, but preserving the dependencies (correlation) between subbands by coupling the likelihoods of adjacent bands [34, 35]. The structure of speech was further exploited by separating the signal into coarse and fine spectral structure (i.e. formants and harmonics/excitation), and having separate, relatively simple models constrain the dynamics of each. This layered model was able to segregate mixed signals, and also to fill-in regions of a signal that were obscured or otherwise missing by propagation of the dynamics [36, 37]. This work was described in detail in the dissertation of Manuel Reyes Gómez [38].

Continuation of this theme of signal separation led to simplified models firstly to simply classify individual time-frequency cells as dominated by a target voice or by noise [41], and to capture the behavior of a speech signal through a simple vector-quantization dictionary, but including phase-derivative information to permit satisfactory resynthesis [7]. This work went on to be the basis of a subsequent NSF project, IIS-0535168: Separating Speech from Speech Noise.

1.1.2 Environmental Sound Classification

While separating individual sources is the most desirable goal, in many circumstances and for many applications it is easier and sufficient to describe and classify the overall result i.e. to recognize the characteristics of the combination. We pursued the automatic recognition and classification of everyday sounds using tools developed for speech recognition such as hidden Markov models (HMMs), addressing the particular issues that arise such as the lack of any natural definition of a "state" (the role served by phonemes in speech recognizers) in non-speech sounds [39]. To investigate the value of these techniques in realistic applications, we studied the problem of analyzing personal audio lifelogs, the kind of continuous recordings of the sounds heard by an individual that are currently quite easy to capture, but very difficult to access. We analyzed this kind of material on a relatively coarse timescale (2. . .60 s segments) to segment into "episodes" consisting of a single activity or acoustic environment, then to cluster together separate instances of similar environments [9, 10, 11]. We collaborated with Microsoft and Dublin City University to study the relative value of acoustic and visual cues in this type of task [2]. To provide more specific, indexible information based on such long-duration audio recordings, we looked at special-purpose detectors to find voices [15], music [16], clap sounds (including estimating distance from the microphone) [20], and exactly-repeating complex sounds such as ringtones or musical excerpts [26].

A second application domain for these techniques is the analysis of the soundtracks of consumer or other environmental video recordings. In collaboration with Kodak, we looked at the problem of classifying short videos shot by digital cameras according to the kinds of conceptual tags that users might want to use in browsing their collections, studying in detail the relative usefulness of audio and video features in recognizing categories such as "beach" or "dancing" [1, 18]. Our most recent work concerns the problem of narrowing down the temporal specificity of these labels from the scope of an entire video clip, to the particular times within that clip that specifically represent the concept [19].

This work is the subject of the dissertation of Keansub Lee [17].

1.1.3 Music Audio Analysis & Classification

Music represents a highly specific yet particularly important class of complex sound mixtures. Drawing on the ideas of complex sound organization developed in the less structured domains described above, we also embarked on a variety of investigations of music audio within this project. Drawing ideas from speech separation, we looked at the task of separating individual notes from music, for instance to transcribe just the predominant melody from complex, polyphonic music, which we treated as a supervised classification task, using labeled pairs of musical spectral slices, and the corresponding melody note to train an SVM classifier [27, 6]. This led to a major formal evaluation involving groups from around the world which we organized and conducted [30]. We also looked at the problem of discriminating between solo and accompanied sections within the music, in part to aid with the automatic collection of training data [40]. We were able to generalize our approach to apply to all the simultaneous notes in polyphonic music for full piano transcription [28], and we investigated various techniques to take advantage of the large amount of unlabeled audio data that could potentially be used to improve classification [29]. This work led to the dissertation of Graham Poliner [31].

A second thread in music audio analysis concerns a broader level of analysis, deciding labels or similarity for music audio taken as a whole. An early project used incremental training of support vector machine (SVM) classifiers to generate a playlist of related music that used "thumbs up"/"thumbs down" style responses (relevance feedback) to refine a classifier for what music the listener wished to hear [25]. This led to the development of larger scale music audio classification, for instance to determine the artist or genre of particular pieces [23]. We also looked at analyzing drum rhythms onto a linear basis of "eigenrhythms" as a way to organize and gain insight into musical styles [8]. Much of this work in basic music audio analysis is summarized in an invited overview paper [3].

A specific application of music similarity is in the identification of "cover songs", alternate interpretations of the same underlying musical piece; our system came top in a formal international evaluation in 2007 [5]. In order to collect more ground-truth human labels at a finer temporal resolution, we created a web-based music labeling game that awarded points based on whether brief tags applied to 10 s excerpts were also used by other players [21, 24]. This allowed us to study and evaluate the problem of propagating labels to different levels of specificity (clip, track, album, artist) via multiple-instance learning [22].

The music audio analysis work started within this project led into a subsequent NSF project, IIS-0713334: Data-Driven Music Audio Understanding.

1.1.4 Marine Mammal Sound Recognition

Another domain that raises many sound organization issues is the underwater world. Through a collaboration with marine biologist Diana Reiss, we became involved in the problem of extracting information about marine mammals from hydrophone recordings. We looked at clustering and characterizing echolocation clicks [12], and at tracking and separating the communication whistles made by dolphins [13]. Because the underwater environment is very noisy and very reverberant, it proved a significant challenge to identify and characterize individual dolphin calls. Our work towards this goal is detailed in the dissertation of Xanadu Halkias [14].

1.2 Education Activities

This project dealt with the core research topics of my lab, and as such fed into all my teaching activities. Over the period of the project, I taught the senior undergraduate/first year graduate Digital Signal Processing course six times, and the graduate-level Speech and Audio Processing and Recognition class six times. Both classes used examples on audio separation and analysis as in-class demonstrations, and drew on the topics of this project as material for the individual class projects required of both classes. Several of these class projects in fact went on to become publications related to this project [8, 10, 20, 27, 40, 26].

During the course of the project I also developed and delivered other classes on Music Content Analysis with Machine Learning, and Music Signal Processing, which drew heavily on the music audio analysis work done in this project. Materials for all these courses are freely available on my web site, as described below.

We also initiated a joint seminar, the Music Engineering Art Project or MEAP, between the engineering school and Columbia's Computer Music Center. This arts-focused collaboration led to many hours of heated discussion and, eventually, some software for music audio analysis and modification, MEAPsoft, which we freely distribute (see below).

1.3 Findings

Each of the research threads mentioned in the activities section resulted in their own findings and conclusions as detailed in the individual publications and summarized in our annual reports. In this section, we try to draw some broad, overall conclusions from the project work:

- Techniques developed primarily for speech recognition such as Gaussian Mixture Models and Hidden Markov Models can be very successful when applied to analogous tasks such as classifying acoustic environments, or recognizing the genre of a piece of music.

- Real-world, unconstrained audio, such as the soundtracks of consumer videos, or commercial music recordings, are extremely dense signals in which it is likely impossible to exactly observe individual sources. We found, however, that many problems could be solved by using features that describe the overall mixture, rather than needing to describe each source individually.

- Support vector machine classifiers, while still relatively rare in speech recognition, almost always give benefits on the kinds of general audio problems we looked at. Both our music genre classification system, and our video soundtrack classification system, used Gaussian models as a first stage, then used SVM classifiers in a representation space derived from the Gaussians; this proved very successful across both domains.

- In extreme situations, such as the very high background noise conditions encountered in consumer video or underwater, special-case solutions can always improve the performance over more generic techniques. We were able to show substantial improvements over standard baselines in voice and music detection for noisy environmental audio, and in dolphin whistle tracking, by developing novel techniques specifically adapted to the target domain.

- In music transcription, supervised classification techniques can compete with traditional methods that explicitly exploit the known structure of musical pitch. Supervised classification has the advantage of embodying fewer prior assumptions about the structure of the signal, and can often be improved without bound simply by employing more training data.

2 Training and development

The project provided full or partial support for the completed Ph.D.s of four former students: Manuel Reyes Gómez, Graham Poliner, Xanadu Halkias, and Keansub Lee. Each of these students was trained in the process of academic research through their involvement in this project. The project also involved shorter-term involvement from three further Ph.D. students (who then went on to be supported by different projects), three MS students, one undergraduate researcher, and three high school summer interns.

3 Outreach

Each of the 26 peer-reviewed conference papers listed in the references was associated with a presentation given by the PI or a student. These occurred at major international conferences such as ICASSP, ISMIR, and ACM Multimedia. In addition, over the course of the project, the PI gave more than 30 invited talks at academic and industrial research labs around the world.

The PI also used the foundation of this project as a basis for organizing or co-organizing six workshops: Statistical and Perceptual Audition SAPA 2004 (Korea), SAPA 2006 (Pittsburgh), and SAPA 2008 (Australia), NSF-sponsored workshops on Separating Speech from Speech Noise (Montreal, 2003, and Montreal, 2004), and Music Information Processing Systems (British Columbia, 2004).

The PI was a guest editor on several related journal issues: Speech Communication – Special Issue on Recognition and Organization of Real-World Sounds (8 papers, Sep 2004), Special Issue of IEEE Transactions on Speech and Audio Processing on Statistical and Perceptual Audio Processing (12 papers, Jan 2006), and Special Issue of IEEE Transactions on Speech and Audio Processing on Music Information Retrieval (20 papers, Jan 2008).

4 Contributions

4.1 Contributions to the principal discipline

Our work in source separation has contributed several novel ideas. We have pushed forward the idea of using source models as constraints for ill-posed single-channel separation, and pioneered the idea of separating voices into excitation and formants to simplify the specification and application of constraints in each domain.

We also pioneered the application of large-scale statistical techniques for segmentation and classification of environmental audio such as audio lifelogs and consumer video soundtracks. We have provided a number of specific, successful demonstrations of how signal processing and machine learning can be deployed to solve user-relevant problems such as locating speech or music, or classifying soundtracks into user-relevant categories.

Our work in music audio analysis has opened up several new areas of research including the general approach of using SVM classifiers for generic music classification, classifier-based music transcription, and music audio tagging/labeling at the level of shorter clips (e.g. 10 s) instead of entire tracks.

4.2 Contributions to sister disciplines

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

In the field of marine mammal audio analysis, we showed how automatic techniques could be used to separate and characterize dolphin calls, something that is currently limited to human labeling. Full automation of this procedure could vastly improve the depth of analysis in the study of dolphin behavior and communication.

4.3 Contributions to human resources

The project has provided direct support for four graduate students, and has enabled the PI to participate in research supervision of numerous other students at graduate, undergraduate, and high school levels.

4.4 Contributions to educational infrastructure

The project has supported the 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 on music audio analysis is motivated by the data management challenges of today’s enormous personal music collections.

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/`
- Presentations from the 2003 Montreal Workshop on Speech Separation `http://labrosa.ee.columbia.edu/Montreal2003/`
- Presentations from the 2004 Montreal Workshop on Speech Separation `http://labrosa.ee.columbia.edu/Montreal2004/`
- Proceedings of the 2004 workshop on Statistical and Perceptual audition: `http://www.sapa2004.org/`
- Proceedings of the 2006 workshop on Statistical and Perceptual audition: `http://www.sapa2006.org/`
- Proceedings of the 2008 workshop on Statistical and Perceptual audition: `http://www.sapa2008.org/`

6 Publications

This project led to four Ph.D. theses (Reyes [38], Poliner [31], Halkias [14], and Lee [17]), nine published journal papers ([3, 25, 6, 13, 11, 4, 30, 28, 24]) with one more currently in revision ([18]), and a further 26 papers published in peer-reviewed international conferences (see references).

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] Aiden R. Doherty, Alan F. Smeaton, Keansub Lee, and Daniel P. W. Ellis. Multimodal segmentation of lifelog data. In David Evans, Sadaoki Furui, and Chantal Soulé-Dupuy, editors, *RIAO*. CID, 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/DohSLE07-lifelog.pdf`.

[3] D. P. W. Ellis. Extracting information from music audio. *Comm. Assoc. Comput. Mach.*, 49(8):32–37, Aug 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/Ellis06-musicinfo-cacm.pdf`.

[4] 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.

[5] 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`.

[6] D. P. W. Ellis and G. E. Poliner. Classification-based melody transcription. *Machine Learning Journal*, 65(2–3):439–456, Dec 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/EllisP06-melody.pdf`.

[7] D. P. W. Ellis and R. J. Weiss. Model-based monaural source separation using a vector-quantized phase-vocoder representation. In *Proc. ICASSP-06*, pages V–957–960, Toulouse, 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/EllisW06-pvocvq.pdf`.

[8] Daniel P. W. Ellis and John Arroyo. Eigenrhythms: Drum pattern basis sets for classification and generation. In *Proc. Int. Symp. on Music Info. Retr. ISMIR-04*, pages 101–106, Barcelona, October 2004. URL `http://www.ee.columbia.edu/~dpwe/pubs/ismir04-eigenrhythm.pdf`.

[9] Daniel P. W. Ellis and Keansub 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. URL `http://www.ee.columbia.edu/~dpwe/pubs/carpe04-minimpact.pdf`.

[10] Daniel P. W. Ellis and Keansub 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*, pages 1–6, Jeju, Korea, October 2004. URL `http://www.ee.columbia.edu/~dpwe/pubs/sapa04-persaud.pdf`.

[11] Daniel P. W. Ellis and Keansub Lee. Accessing minimal-impact personal audio archives. *IEEE MultiMedia*, 13(4):30–38, Oct-Dec 2006. doi: 10.1109/MMUL.2006.75. URL `http://www.ee.columbia.edu/~dpwe/pubs/EllisL06-persaud.pdf`.

[12] 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. URL `http://www.ee.columbia.edu/~dpwe/pubs/HalkE06-clicks.pdf`.

[13] X. Halkias and D. P. W. Ellis. Call detection and extraction using Bayesian inference. *Applied Acoustics*, 67(11-12):1164–1174, Nov-Dec 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/HalkE06-whistles-aa.pdf`.

[14] Xanadu C. Halkias. *Detection and Tracking of Dolphin Vocalizations*. PhD thesis, Columbia Univ. Dept. Elec. Eng., 2008.

[15] 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`.

[16] K. Lee and D. P. W. Ellis. Detecting music in ambient audio by long-window autocorrelation. In *Proc. ICASSP*, pages 9–12, Las Vegas, Apr 2008. URL `http://www.ee.columbia.edu/~dpwe/pubs/LeeE08-musicdet.pdf`.

[17] Keansub Lee. *Analysis of Environmental Sounds*. PhD thesis, Columbia Univ. Dept. Elec. Eng., 2009.

[18] Keansub Lee and D. P. W. Ellis. Audio-based semantic concept classification for consumer video. *IEEE Tr. Audio, Speech, Lang. Proc.*, (submitted), 2009.

[19] Keansub Lee and D. P. W. Ellis. Detecting local semantic concepts in environmental sounds using markov model based clustering. In *Proc. ACM MultiMedia*, page (submitted), Beijing, Oct 2009.

[20] Nathan Lesser and Daniel P. W. Ellis. Clap detection and discrimination for rhythm therapy. In *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Proc.*, pages III–37–40, Philadelphia PA, March 2005. URL `http://www.ee.columbia.edu/~dpwe/pubs/icassp05-claps.pdf`.

[21] M. Mandel and D. Ellis. A web-based game for collecting music metadata. In *Proc. Int. Conf. on Music Info. Retr. ISMIR-07*, pages 365–366, Vienna, Austria, 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/MandE07-majorminer.pdf`.

[22] M. Mandel and D. P. W. Ellis. Multiple-instance learning for music information retrieval. In *Proc. ISMIR*, pages 577–582, Philadelphia PA, Sep 2008. URL `http://www.ee.columbia.edu/~dpwe/pubs/MandelE08-MImusic.pdf`.

[23] M. I. 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*, pages 594–599, London, Sep 2005. URL `http://www.ee.columbia.edu/~dpwe/pubs/ismir05-svm.pdf`.

[24] M. I. Mandel and D. P. W. Ellis. A web-based game for collecting music metadata. *J. New Music Research*, 37(2):151–165, 2008. URL `http://www.ee.columbia.edu/~dpwe/pubs/MandelE08-majorminer.pdf`.

[25] M. I. Mandel, G. E. Poliner, and D. P. W. Ellis. Support vector machine active learning for music retrieval. *ACM Multimedia Systems Journal*, 12(1): 3–13, Aug 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/MandPE06-svm.pdf`.

[26] 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`.

[27] G. Poliner and D. P. W. Ellis. A classification approach to melody transcription. In *Proc. International Conference on Music Information Retrieval ISMIR*, pages 161–166, London, Sep 2005. URL `http://www.ee.columbia.edu/~dpwe/pubs/ismir05-svm.pdf`.

[28] G. Poliner and D. P. W. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 2007 (2007):9 pages, 2007. doi: 10.1155/2007/48317. URL `http://www.ee.columbia.edu/~dpwe/pubs/PoliE06-piano.pdf`. Special Issue on Music Signal Processing.

[29] G. Poliner and D. P. W. Ellis. Improving generalization for polyphonic piano transcription. In *Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous.*, pages 86–89, Mohonk NY, Oct 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/PoliE07-semisup.pdf`.

[30] 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.*, 15(4):1247–1256, May 2007. doi: 10.1109/TASL.2006.889797. URL `http://www.ee.columbia.edu/~dpwe/pubs/PolEFGSO07-meleval.pdf`.

[31] Graham E. Poliner. *Classification-Based Music Transcription*. PhD thesis, Columbia Univ. Dept. Elec. Eng., 2008.

[32] Manuel Reyes-Gomez, Bhiksha Raj, and Daniel P. W. Ellis. Multi-channel source separation by beamforming trained with factorial hmms. In *Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous.*, pages 13–16, Mohonk NY, October 2003. URL `http://www.ee.columbia.edu/~dpwe/pubs/waspaa03-muchan.pdf`.

[33] Manuel Reyes-Gomez, Bhiksha Raj, and Daniel P. W. Ellis. Multichannel source separation by factorial hmms. In *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Proc.*, pages I–664–667, Hong Kong, April 2003. URL `http://www.ee.columbia.edu/~dpwe/pubs/waspaa03-muchan.pdf`.

[34] Manuel Reyes-Gomez, Daniel P. W. Ellis, and Nebojsa Jojic. Multiband audio modeling for single channel acoustic source separation. In *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Proc.*, pages V–641–644, Montreal, 2004. URL `http://www.ee.columbia.edu/~dpwe/pubs/icassp04-muband.pdf`.

[35] Manuel Reyes-Gomez, Nebojsa Jojic, and Daniel P. W. Ellis. Detailed graphical models for source separation and missing data interpolation in audio. In *Proc. Snowbird Learning Workshop*, Snowbird, April 2004. URL `http://www.ee.columbia.edu/~dpwe/pubs/snowbird04-transform.pdf`.

[36] Manuel Reyes-Gomez, Nebojsa Jojic, and Daniel P. W. Ellis. Towards single-channel unsupervised source separation of speech mixtures: The layered harmonics/formants separation-tracking model. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04*, pages 37–42, Jeju, Korea, October 2004. URL `http://www.ee.columbia.edu/~dpwe/pubs/sapa04-transform.pdf`.

[37] Manuel Reyes-Gomez, Nebojsa Jojic, and Daniel P. W. Ellis. Deformable spectrograms. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 285–292, Barbados, Jan 2005. URL `http://www.ee.columbia.edu/~dpwe/pubs/aistats05-defspec.pdf`.

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

[39] M.J. Reyes-Gomez and D.P.W. 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*, pages I–73–76, Baltimore, July 2003. URL `http://www.ee.columbia.edu/~dpwe/pubs/icme03-hmmsel.pdf`.

[40] C. Smit and D. P. W. Ellis. Solo voice detection via optimal cancelation. In *Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous.*, pages 207–210, Mohonk, NY, Oct 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/SmitE07-solo.pdf`.

[41] 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`.