Separating Speech from Speech Noise
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 29, 2007

1 Activities & Findings

1.1 Research Activities

The main work at Columbia this year has been the development of algorithms for extracting and recognizing speech in nonstationary, noisy environments when only a single microphone channel is available. Our particular approach is based on using trained models to distinguish regions of time-frequency containing speech from nonspeech areas [2], and we have pursued this along several directions: One approach is to use trained models of the speech signal and to find the best set of model parameters that are consistent with the noisy speech observations. An alternative approach is to treat the labeling of each time-frequency cell as a simple classification task, and to train pattern recognition classifiers to perform this task. In that work, the challenge is to find the best classifier architecture and the most effective representation of the context.

When more than one microphone channel is available, some different approaches to source separation become possible. The two-channel case is particularly interesting because this is the number of ears possessed by the typical listener. We have been looking at ways to separate sources in recordings made by dummy-head microphones (which attempt to closely duplicate the signals at the ears of a listener), and in particular we have looked at the case of multiple, overlapping speech sources in reverberant environments. Reverberant reflections from walls etc. make the between-channel cues to source direction much less well defined, requiring a more complex, probabilistic model of the observations. We use the well-known EM algorithm to converge to a good local solution.

At a more theoretical level, we have been developing the dual of the well-known linear prediction model. In its frequency-domain dual, a linear predictor of values in the frequency domain results in a parametric description of the temporal envelope – i.e. the Hilbert magnitude. This novel theoretical tool can be useful as a way of finding features in complex sounds, and for interpolating or reconstructing partial observations.

In support of our interest in statistical models of the entire speech signal, we have also been looking at voice pitch, and the extent to which simple, stylized models of voice pitch still carry the perceptually-important information. To measure this, we have constructed an experiment in which our simple pitch model, which fits only a linear pitch contour to each syllable (for about 10 dimensions per second) is compared to the original pitch contour by listeners to see how easily they can tell the difference. Both pitch contours are resynthesized through the same high-quality synthesis engine [4].

Finally, in support of our collaborators within the project who are developing perceptual experiments, we constructed a special-purpose derivative of a speech recognizer specifically for making automatic phone boundary labels of known utterances. Most speech recognizers operate at a 10 ms frame rate, since speech information is generally consistent on that time scale, and little is gained in recognition accuracy by using a finer resolution. However, for the purpose of editing and aligning speech tokens to create controlled, modified speech to be used in perceptual experiments, 10 ms is too coarse a resolution. Hence, we trained a new set of acoustic models at a 1 ms resolution, and packaged them up in a system that takes a voice recording and a transcript of the words spoken, then makes a high-resolution time alignment between the best-fitting sequence of words for that utterance and the observed signal. This has been used by collaborators at EBIRE and Boston to construct experiments in which information is carefully interchanged between similar but different speech recordings to better understand exactly which part of the signal is most critical to the perceived information.

1.2 Educational activities

The educational activities arising from this project include:

- Supervision of the Columbia graduate student, Ron Weiss, who is fully supported by this project. In addition to research and development of speech separation models, Ron also made his first oral presentation at an international conference this year [7].

- Partially-supported student Mike Mandel was also supervised and made two presentations, one at the highly prestigious NIPS conference [5, 6].

- The work on pitch modeling and perception was performed by an undergraduate researcher, Suman Ravuri. We hope to publish this work in the coming year.

- My lab hosted a visit by Adam Lammert, the research scientist working on this project at EBIRE, who came to New York for a week to describe his work, and to learn more about ours. This very productive visit allowed him to return to EBIRE and make full use of the speech alignment tool we had developed. He has even enhanced to tool to handle longer sentences.

- The research also fed directly into my graduate class on Speech and Audio Processing and Recognition (ELEN E6820), which includes modules on speech perception, and on signal separation, in which I was able to describe and demonstrate some of the ongoing work in this project. The class involves a large individual semester project by each student, and several projects have been based on and inspired by this work.

Presentations made by the PI that included material from this project included an invited paper at ICASSP in Toulouse [3], an invited talk at the Ecole Normale Superieure in Paris on speech separation, and an invited talk at the NIPS Workshop on model-based sound separation.

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 [3]. The "good" time-frequency cells can be identified with separate models trained to recognize the local characteristics of unmasked speech components [7].

- 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 [5]. The EM algorithm can be usefully employed to identify a small number of source directions, and to segment the sound signal according to those source directions, even in the presence of reverberation and noise [6].

- Linear prediction is a very useful and powerful model even when applied directly to frequency-domain representations such as the DCT, rather than the conventional application to time-domain waveforms. The parametric models for signal envelopes can be used for coding, resynthesis, and recognition [1].

- A very simple model of voice pitch, which fits only a single linear segment to each spoken syllable, provides highly natural sounding resynthesis i.e. it appears that listeners are largely insensitive to the fine details of pitch variation visible in pitch tracker outputs. (This work is still being prepared for publication).

- Adjusting the frame rate of a conventional speech recognizer to a much finer time base can successfully be used to make automatic alignments of signals to know pronunciations with a resolution of 1 ms (or less, if required).

2 Training and development

This year, the project has provided full or partial support for three graduate students: Ron Weiss, Michael Mandel, and Marios Athineos, 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.

We have also worked with undergraduate Suman Ravuri who gained his first experience in both developing a novel algorithm, and in designing and conducting perceptual experiments to verify its capabilities.

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.

4 Contributions

4.1 Contributions to the principal discipline

We have presented and described ideas and algorithms for the separation of sound sources in real, complex environments based on their intrinsic properties as well as their spatial characteristics.

4.2 Contributions to sister disciplines

Through the adaptation of speech recognition techniques to provide automatic, high-resolution, and consistent labels to speech tokens, we have made possible large-scale, methodologically sound experiments in the perception of speech stimuli by listeners. These experiments are in process at collaborators BU and EBIRE.

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 several more both in and out of class.

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

The motivation behind the speech separation project is to develop and verify techniques for separating speech that have real benefits for human listeners, both normal and hearing impaired. We are discussing this project with several hearing aid manufacturers, since they might be well paced to bring the benefits of this work to the marketplace.

5 Resources made available

Resources relating to this project are made available online at `http://labrosa.ee.columbia.edu/speechsep/`. The include:

- Examples of real-world recordings posing problems to speech separation.
- Package of code for aligning recordings to phone labels with very high resolution.
- Class materials (slides, assignments, practicals, demonstrations) for Speech and Audio Processing and Recognition: `http://www.ee.columbia.edu/~dpwe/e6820/`
- Proceedings of the 2006 workshop on Statistical and Perceptual audition: `http://www.sapa2006.org/`

6 Publications

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

References

[1] M. Athineos and D. P. W. Ellis. Autoregressive modeling of temporal envelopes. *IEEE Tr. Signal Proc.*, accepted, 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/AthinE07-fdlp.pdf`.

[2] D. P. W. Ellis. Model-based scene analysis. In D. Wang and G. Brown, editors, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, chapter 4, pages 115–146. Wiley/IEEE Press, 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/Ellis06-casamodels.pdf`.

[3] 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`.

[4] H. Kawahara, I. Masuda-Kastuse, and A. de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication. 27, 187–207.*, 27:187–207, 1999.

[5] 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`.

[6] 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`.

[7] 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`.