Separating Speech from Speech Noise to Improve Intelligibility

Since the expression 'cocktail party problem' was coined by Cherry [1953], much progress has been made in sound source segregation algorithms. Despite some impressive results, these algorithms are rarely used in hearing aids, hands-free telephone sets, etc. The reason is that none of these algorithms can guarantee separation in all circumstances. Failure degrades intelligibility, all the more so as processing artifacts may masquerade as speech sounds and mislead the listener. In addition to developing the basic separation algorithms, we will build upon them from two new angles: First, models of speech (derived from machine learning techniques) are used to *interpolate* over missing portions of the signal and resynthesize speech to fill gaps where separation has failed. Second, the listeners' ability to use internal models to *perceptually restore* the missing speech is enhanced by adding carefully crafted noise to mask missing spectrotemporal regions. The two approaches may be combined according to a tradeoff between erroneous interpolation on one hand, and cognitive effort on the other. Our team's expertise spans computer science, engineering, and psychology to address the following subproblems:

- Characterization of the statistics of speech in noisy surroundings, and failure modes of current source separation algorithms. These are needed to quantify the distribution over time and spectrum of the cues that segregation algorithms require, and to quantify the spectro-temporal distribution of missing features.
- Interpolation and synthesis using dynamic speech models based on machine-learning techniques. It is essential to quantify *reliability* of this interpolation.
- Psychoacoustics of the perceptual restoration of interrupted speech by the addition of noise. Existing results need to be complemented in a systematic way to specify the spectrotemporal characteristics that best favor restoration and best minimize interference, and to provide results that are most useful in guiding algorithm development.
- Integration and evaluation. Issues are how to select the best segregation strategy, when to provide the listener with interpolated speech or noise to promote perceptual restoration, and how to measure the performance with listeners in realistic conditions.

Intellectual merit: This project will advance signal separation techniques based on machine learning and human processing models. Unlike previous work on blind source separation, the specification of a clear, high-level, and application-relevant goal – improving intelligibility for human listeners – provides a firm grounding and highlights new challenges. These include developing an understanding of how humans comprehend and restore speech in a background of nonstationary noise and how best to exploit the imperfect analysis and separation of mixtures to recreate sounds best suited to a listener.

Broader impact: Successful techniques for speech enhancement in situations with realistic, dynamic noise have far-reaching applications including hearing prostheses, recovery of archival recordings, and various security/forensic areas. Moreover, by cementing a nascent collaboration between technologists and psychologists, this project establishes the disciplinary links required for a complete solution to speech separation. The project will involve educational initiatives at the participating universities, and interdisciplinary exchange of the supported students among labs in U.S. and France.

Separating Speech from Speech Noise to Improve Intelligibility

Daniel P.W. Ellis,
Alain de Cheveigné, Pierre L. Divenyi, Barbara Shinn-Cunningham, DeLiang Wang

1 Introduction

Cherry [1953] introduced the term "cocktail party problem" to refer to the ability of human listeners to follow and understand a single voice amidst high levels of interference, and it has been widely used as a synonym for blind source separation – the problem of 'unmixing' a collection of acoustic signals that have been combined in a single medium. Of course, human listeners never actually separate signals, but instead recover the messages and information about each voice from the mixture. Listeners accomplish this by using signal features as well as knowledge and constraints about what the signal and message may be. These constraints arise at many different levels, from an implicit understanding of acoustics (sounds emanating from a vocal tract, or coming from a particular direction, have certain consistent properties) to very abstract assumptions about meaning and content.

This project aims to unite the technological and psychological aspects of the cocktail party problem by investigating in detail how listeners can understand speech in realistic situations such as crowded rooms, and by developing novel algorithms and signal processing systems that can actually improve listeners' understanding of such speech signals, for instance from recordings made of lively meetings.

A wide variety of signal separation algorithms have been proposed, based on input from single or multiple microphones. They range from acoustic beamforming [Griffiths and Jim, 1982] and Independent Component Analysis (ICA) [Bell and Sejnowski, 1995] to Computational Auditory Scene Analysis (CASA) [Cooke and Ellis, 2001, Brown and Wang, 2005] and statistical model-based methods [Roweis, 2003].

None of these algorithms, however, is able to do a complete job of separating signals in the kind of noisy, chaotic circumstances that humans cope with. Rather than focusing on incremental improvements of the algorithms in the hope that they will eventually conquer all challenges, this project recognizes the inevitability of imperfect separation and considers how best to cope with it. In the face of imperfect separation, what is the best way to present signals that delivers maximum value (i.e., intelligibility) to a human listener? This casts an interesting new light on the problem: it forces us to consider how best to convey to the listener the *partial information* that may be salvaged in these situations – since partial information is often sufficient for understanding. The speech code is 'redundant' as revealed by early experiments showing that no part of the speech spectrum is 'irreplaceable': intelligibility remains high regardless of which frequency regions are filtered out, provided that enough remain [French and Steinberg, 1947]. The same applies to the time dimension: short segments of speech can be deleted with little perceptual degradation of the speech message [Miller and Licklider, 1950, Warren, 1970], pointing to the robust nature of human speech communication. Automatic recognition systems have a long way to go in duplicating this level of robustness, although recent work on 'missing data' speech recognition

shows promise [Cooke et al., 2001, Barker et al., 2005]. Redundancy implies that imperfect segregation is not fatal if the partial information recovered is sufficient to determine the message – provided it can be effectively conveyed to the listener.

So what could automatic algorithms do to help human listeners deal with incomplete speech? When the source separation algorithm fails, the machine can either (a) infer the missing parts of the signal and insert synthetic substitutes to create an apparently complete signal (whose reliability depends on the inferential models applied), or (b) produce a new 'noisy speech' signal composed of the reliably separated portions of the target, combined with a "minimally informative" noise that conveys some limited information about the missing portions, or at least minimizes the disruption caused by leaving silent 'gaps' in the recovered speech, since filling gaps with noise is known to potentially improve intelligibility [Miller and Licklider, 1950, Warren, 1970]. How best to design this replacement noise, and how to automatically choose between synthesizing interpolated speech or replacement noise, are entirely new problems to be addressed in this project. We consider the novel problem of presenting a "best guess" reconstructed signal to a listener that *includes* information about the certainty of the reconstruction. This kind of processing is only possible when speech is enhanced through a combination of complex signal models and signal separation techniques.

This project is unique in its combination of engineering and computer science with experimental psychology, involving an equal exchange between these two broad areas. It is quite common for engineers to take perceptual phenomena as a starting point; similarly, experimental psychologists often develop computer models to explain and illuminate their observations and to generate testable predictions. In this project, however, leading labs in both psychology and engineering are brought together to focus on the process of building signal separation systems that are based upon, and guided by, perceptual results. In this way, we hope both to deepen our understanding of the perception of speech in noise, and to enhance our capability to design systems to perform similar tasks. These will be based on novel algorithms whose principal goal is to afford an intelligibility gain for human listeners under challenging, real-world conditions.

The most similar work of which we are aware is in the field of hearing aids. Hearing aids target only impaired listeners, and are subject to constraints such as limited power, size, and cost. These factors tend to restrict the range of processing schemes considered, and current hearing aids have a poor track record in ensuring intelligibility in noisy environments. We wish to target a wider range of potential users, including normal listeners, and follow a much more ambitious approach to separating and enhancing speech. This will lead to a different flavor of solution, which should, in due course, be applicable to hearing prostheses.

Compared to prior work in signal separation, our emphasis on improving *intelligibility* is distinctive: most signal separation work targets improvement on relatively simple objective measures of performance such as signal to noise ratio (SNR) (although see e.g. Kollmeier and Koch [1994]). This is an outstanding problem in source separation, since the nonlinear processing involved often cannot be well characterized with simple measures like SNR [Ellis, 2004]. Here we circumvent this problem by specifying intelligibility as the goal – an application-relevant achievement that we can, with sufficient care, measure directly.

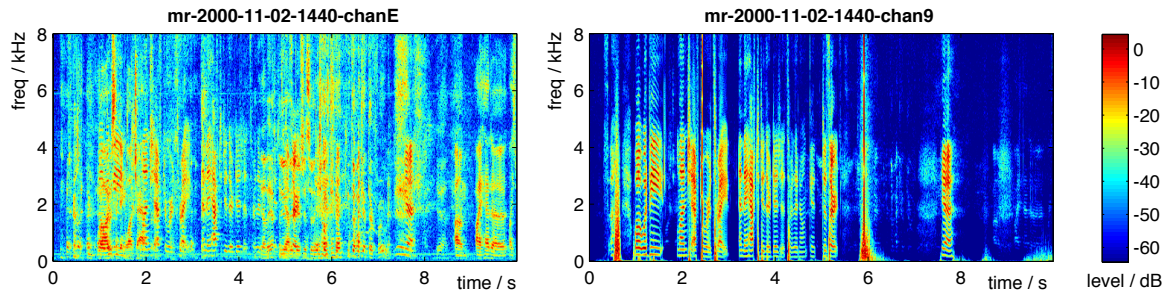


Figure 1: Spectrogram of real-world recording of voices in a meeting. Left: tabletop mic; right: head-mounted mic for one participant, whose speech is partly unintelligible due to overlap from other speakers in the tabletop recording. Our goal is to improve intelligibility in such situations.

Figure 1 shows an example of the noisy situations we are considering, the kind of acoustic challenge that we are exposed to daily. It shows the spectrogram of a few seconds from a recording made by a 'personal recorder' (a small flash memory device with a microphone; this and other examples can be heard at `http://labrosa.ee.columbia.edu/speechsep/`). Normal listeners with young ears and intact cognitive resources may be able to make sense of the speech embedded in the babble, but those who are hearing impaired, or simply elderly, or engaged in tasks that tax their attention, may have more difficulty.

Interference of the kind shown in fig. 1 degrades speech in two ways: First, the level of noise may exceed the dynamic range of auditory processing within each channel and mask evidence of the speech – so-called *energetic masking*. Second, portions of interference may be mistaken for or confused with parts of the target speech, causing what is known as *informational masking*. Intelligibility can thus be degraded even when both sources are perfectly audible (i.e. energetic masking has no impact). For instance, when listening to competing speech signals listeners may hear and understand both talkers but get confused about which talker pronounced a word [Brungart and Simpson, 2001, Arbogast et al., 2002, Shinn-Cunningham et al., 2004]. This form of interference appears to reflect an inability to ignore a competing source that is similar to the source of interest [Durlach et al., 2003]. Masking by speech is greater, for instance, than by time-reversed speech or noise [Hawley et al., 2004], and can actually be reduced by adding still more talkers to a mixture [Brungart and Simpson, 2001]. Informational masking is reduced when sources differ in ways that allow a listener to direct their attention and resolve competition between similar sources. Attention can be anchored on various acoustic or perceived features, such as a talker's voice [Darwin and Hukin, 2000], or perceived spatial separation between sources [Brungart and Simpson, 2002] (which is little affected by reverberation [Shinn-Cunningham, 2002, Kidd et al., 2005]).

Our emphasis in this project is basic science and development of fundamental algorithms. At the same time, since we will focus on real-world noise and recordings, our results will apply immediately to meaningful everyday situations. Review of personal audio recordings (including 'lifelog' archives [Ellis and Lee, 2004b] as well as recorded meeting [Morgan et al., 2001]) is an important practical application for these techniques, particularly

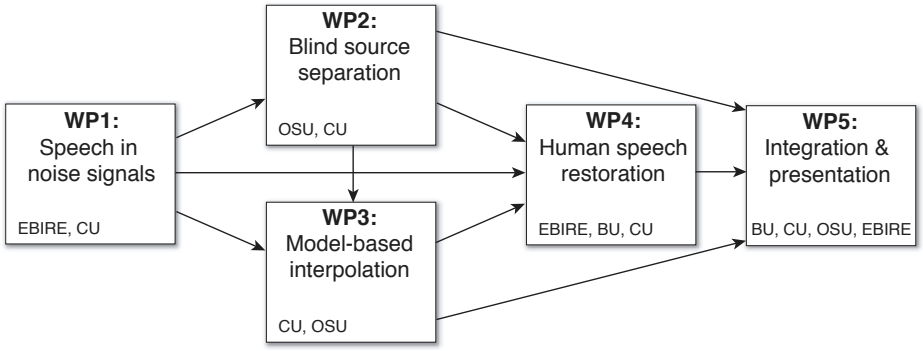


Figure 2: Workpackages for the proposed research and their interrelations. Also shown are the main sites involved in each task.

since the processing can be offline (non-realtime). Because our approaches do not rely on the precise geometry of multi-microphone sensors, a wide range of existing recordings will be amenable to these approaches. Although we will consider only one- and two-channel recordings, our techniques have natural extensions to multi-microphone settings. We focus on speech noise (i.e. competing talkers) not to simplify the problem but to ensure that we address the most challenging case of highly dynamic and confusable interference; speech noise also provides a useful continuum to more uniform noise as more voices are added. We do restrict ourselves, however, to algorithms that can eventually be implemented in real-time systems, enabling a new generation of high-performance hearing aids that would be valuable even to normal-hearing subjects in high-noise environments.

2 Proposed Research

Figure 2 presents an overview of our proposed research. We have divided the project into five 'workpackages', each dealing with a separate aspect of the overall problem. Some consist primarily of psychoacoustics, some of engineering/computer science, and some integrate the two. After a brief description of each work package and how they fit together, the remaining parts of this section describe the content of each workpackage at greater length.

Workpackage 1 begins by examining in detail the kinds of signals that we wish to handle. The domain of "speech in noise" covers a very wide range of signals, and we limit ourselves to realistic situations where the interference is principally other voices, though varying in number, distance, and under varying room conditions (reverberation). Within that space, we wish to deepen existing understanding of the properties of the target, for instance the distribution of local signal-to-noise ratios (SNR) under different time-frequency decompositions, and the size and distribution of 'islands' of favorable SNR that appear to mediate speech according to the glimpsing model [Cooke, 2005]. This workpackage will also address the creation of realistic synthetic target-in-noise signals, validated by comparison to recorded data.

Workpackage 2 is concerned with the separation of signals. Its scope includes beamforming and ICA algorithms, although the main focus will be on the highly nonlinear forms of signal separation based on CASA [Cooke and Ellis, 2001, Brown and Wang, 2005]. In addition to extending our current work to make such separation systems more effective, we will also seek to characterize when and how these approaches fail.

Workpackage 3 complements WP2 by looking at techniques that model the source signal to infer the acoustic waveform of the target over parts of speech where segregation algorithms fail based on separated portions. The models will be derived from those used in speech coding and speech recognition. Models may also be used to assist segregation by identifying regions of signal that likely come from the target and excluding time-frequency cells that do not fit (as is done explicitly in the "speech fragment decoder" [Barker et al., 2005]). Once again, we are also interested in when and how these algorithms fail.

Workpackage 4 is primarily concerned with the perception of speech degraded by noise or imperfect processing. Given the characterization of speech in noise from WP1, and the failure patterns of segregation algorithms in WP2 and WP3, how well can listeners cope with the imperfectly segregated speech? To what degree may their performance be improved by adjusting the output of the segregation algorithms, for example by replacing gaps by non-informative noise or stimuli crafted to imply the uncertain inferences of models?

Workpackage 5 ties together these results to generate signals that provide the listener with improved intelligibility, given all exploitable sources of knowledge and all available segregation techniques. This includes arbitrating between the multiple signal separation schemes of WP2 and WP3, and between specific restored signals from WP3 or the 'proxy noise' developed in WP4 (or, in general, a continuum between the two). The design of this reconstructed signal is tightly dependent on its effectiveness with listeners, so this WP includes the psychoacoustic evaluation and diagnosis of the overall processing, and will involve close collaboration between all the partners in the project.

2.1 WP1: Statistics of Speech in Noise

Our aim is to get better knowledge of statistics of speech in noise, in particular the distribution of segregation cues and local SNR over the time-frequency plane. Together with known properties of the segregation algorithm, this determines the distribution of missing information in the recovered speech. For example, a beamforming algorithm will fail if interference comes from the same direction as the target. A CASA algorithm based on harmonicity will fail where target and interference are not voiced, or both have the same fundamental frequency. Any algorithm will fail where the local SNR is too low for its dynamic range. Conversely, islands of high SNR may emerge without the need for processing. We need to know the extent and distribution of such islands and gaps.

To gather the required statistics, we will employ our ongoing work with 'personal audio' recordings using sets of body-worn microphones [Digital Innovations, 2003] over extended periods of time in 'day-to-day' environments [Ellis and Lee, 2004b]. These recordings may then be used to calculate statistics of segregation cues (harmonicity, intermicrophone cor-

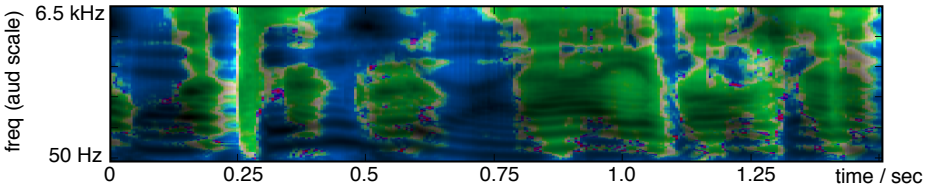


Figure 3: Spectrogram of a mixture of two voices showing regions where the local SNR is dominated by either the first (green) or the second (blue) [Cooke, 2004].

relation) and SNR as they occur in natural speech-in-noise situations, and also to calibrate more controlled simulations using databases and mixing models. Details are omitted for reasons of space, but developing virtual acoustics methods [Gardner, 1998] to synthesize highly accurate replicas of real-world noisy speech is crucial to allow flexible, synthetic test mixtures for use in psychological experiments to ensure that their results accurately reflect real-world performance.

The statistics of speech in noise are used to predict the spectrotemporal distribution of 'segregation failures' of the algorithm under consideration. In case of failure, an algorithm may either allow feed-through from the interference or else delete both target and interference. This produces a 'spectrotemporal map' of gaps and islands. For psychoacoustic experimentation (section 2.4), the map can be applied directly to mixtures by either deleting spectrotemporal zones or replacing them with noise.

Figure 3 illustrates a map of local SNRs within a mixture of two voices from Cooke [2004]. An algorithm capable of correctly assigning the time-frequency pixels dominated by either voice [Hu and Wang, 2004, Roweis, 2001, Bach and Jordan, 2004] will yield a representation with gaps corresponding to pixels dominated by the other voice. Clean speech is 'sparse' in the time-frequency plane, and thus most cells of a mixture are dominated by one or the other voice [Roweis, 2003, Rickard and Yilmaz, 2002]. To what degree this is true for real-world mixtures remains to be determined. It is also important to know the statistics of segregation cues (harmonicity, cross-microphone correlation) within real environments to characterize the usefulness of algorithms relying on these cues.

2.2 WP2: CASA and Blind Source Separation

Speech possesses a rich set of structural features that can be used to perform blind source separation. Psychophysical studies on auditory scene analysis (ASA) show that pitch (harmonicity), common amplitude modulation, and common onset are the main features for human sound organization [Bregman, 1990]. This workpackage will study these ASA features to achieve speech separation. More specifically, we will investigate the use of monaural cues of pitch, onset, and common amplitude modulation to perform simultaneous grouping, and the binaural cues of interaural time difference (ITD) and interaural intensity difference (IID) to perform sequential grouping. This strategy is motivated by psychoacoustic observations that binaural cues may not play a dominant role in simultaneous grouping,

but spatial locations can strongly influence sequential grouping [Darwin and Hukin, 1997, 1999].

An input mixture will be first processed by an auditory periphery model that consists of cochlear filtering [Patterson et al., 1988] and hair cell transduction [Meddis, 1988]. This model yields a cochleagram that is a two-dimensional time-frequency (T-F) representation. The basic element in the cochleagram is a T-F unit. For periodicity analysis we will compute autocorrelation within each T-F unit to add a third dimension, creating a correlogram [Slaney and Lyon, 1992]. Multipitch tracking will produce multiple overlapping pitch contours [Wu et al., 2003, de Cheveigné and Baskind, 2003], and T-F units will be labeled according to their dominant contour. To correctly assign wideband high-frequency channels to pitch contours, we will perform amplitude modulation analysis and label units accordingly [Hu and Wang, 2004]. In addition to recovering a target source from the mixture, techniques based on cancellation of strong periodic intrusions will be investigated; this analysis can be particularly effective for dealing with speech noise [de Cheveigné, 2004].

The end result of the above analysis will be a sequence of short fragments, often only one syllable long. Sequential organization based on sound location will group these across intermittent sections of unvoicing, silence, or other interference. We will first extract interaural time and intensity differences (ITD and IID) within corresponding T-F units of the two channels. ITDs are more effective in the low-frequency range whereas IIDs are more effective in the high-frequency range. Since these values will operate over entire fragments instead of single T-F units, the estimates will be robust to noise. Sequential organization will form continuous streams that have similar between-channel characteristics.

We will also explore spatial filtering (or beamforming) and ICA techniques for speech separation [van Veen and Buckley, 1988, Krim and Viberg, 1996, Hyvärinen et al., 2001]. For adaptive beamforming with two sensors, one can cancel a single noise source. We will apply beamforming in individual frequency channels leading to a subband version [Liu et al., 2001] which could cancel multiple intrusions. One can also extend two-sensor ICA to handle more than two sources – the overcomplete case [Lee et al., 1999, Winter et al., 2004].

Speech separation becomes substantially more complex in presence of room reverberation which is an inescapable factor in real world situations. One approach is to first achieve sound localization in reverberant conditions by modeling the precedence effect, whereby the auditory system fuses the first wavefront and its echoes to localize a sound in space, then design a binaural adaptive filter to retain only sound energy from a specified location. A promising method is to devise a binary mask that retains T-F units where the energy from the specified location is stronger than the remaining energy [Palomäki et al., 2004]. Monaural dereverberation techniques, including recent systems based on machine learning to estimate inverse filters [Nakatani and Miyoshi, 2003, Wu and Wang, 2005], will also be employed. We will first determine the extent to which monaural grouping cues, such as pitch and onset, are disrupted by reverberation, and develop monaural separation methods that are robust in reverberant conditions.

Unvoiced speech lacks the harmonicity cue and, taken alone, is noise-like. Also, un-

voiced consonants are more susceptible to reverberation effects than voiced speech. Despite recent efforts on unvoiced speech segregation [Hu and Wang, 2003, 2005], such signals in the presence of strong interference or reverberation will likely create obscured gaps in separated speech. Another point of likely failure occurs when multiple pitch tracks cross each other in time. In this situation, pitch tracking will be unreliable and pitch-based grouping tends to label T-F segments incorrectly. We will systematically evaluate and characterize these failure modes in relation to the noisy signal statistics from WP1.

2.3 WP3: Model-Based Separation and Interpolation

The signal-based approaches of the previous section are limited to what can be directly recovered from the (suitably transformed) signal. The speech signal, however, incorporates a great deal of redundancy and predictability, and additional separation can be achieved by exploiting this structure. The core idea is that a partial or corrupt observation of a speech signal can be matched to a model that incorporates the intrinsic constraints of a plausible speech signal, and that model in turn provides sufficient constraints to make accurate predictions of the unobserved aspects. One way to represent this model is as a set or summary of a collection of training samples that captures the universe of allowable 'states' or observations of a source signal. Denote signal observations as $\mathbf{x}$, partitioned into known ($\mathbf{x}_k$) and unknown ($\mathbf{x}_u$) dimensions, and assume we can learn the distributions of the observations under model states $p(\mathbf{x}|S)$, where S indexes states that are sufficiently fine-grained to allow the separate dimensions to be approximated as conditionally independent given the state i.e. $p(\mathbf{x}|S) \approx p(\mathbf{x}_k|S)p(\mathbf{x}_u|S)$. Then the distribution of the unknown components may be inferred as [Cooke et al., 2001]:

$$p(\mathbf{x}_u|\mathbf{x}_k) = \sum_S p(\mathbf{x}_u|S)p(S|\mathbf{x}_k) = \frac{\sum_S p(\mathbf{x}_u|S)p(\mathbf{x}_k|S)p(S)}{p(\mathbf{x}_k)} \quad (1)$$

This relation becomes most useful when the discrimination of model states given known components for the model states makes a single factor $p(\mathbf{x}_k|S)p(S)$ dominate in the summation over S. The issues for this model include the granularity of S, the scope of $\mathbf{x}$ (i.e. the temporal window of modeling), and thus whether the inference will result in a tight-enough distribution for $\mathbf{x}_u$ to permit reconstruction of a single value (e.g. the mode of the distribution), or merely specification of a likely range [Barker et al., 2005].

In the implementation of Roweis [2003], the signal model is a vector-quantizer (VQ) that memorizes representative time-frequency patches of training speech. To recover speech that, for instance, has a known range of frequencies missing or obscured, the VQ codebook is searched for matches across the known dimensions; the best match then gives estimates for the remainder of the spectrum simply by using the rest of the codeword.

When the missing components are not known *a priori*, the division into $\mathbf{x}_k$ and $\mathbf{x}_u$ must be achieved by other means – the problem of estimating the time-frequency mask for a particular source discussed in section 2.2 above. Barker et al. [2005] address this problem while simultaneously inferring the states S (which leads to identifying the words in the

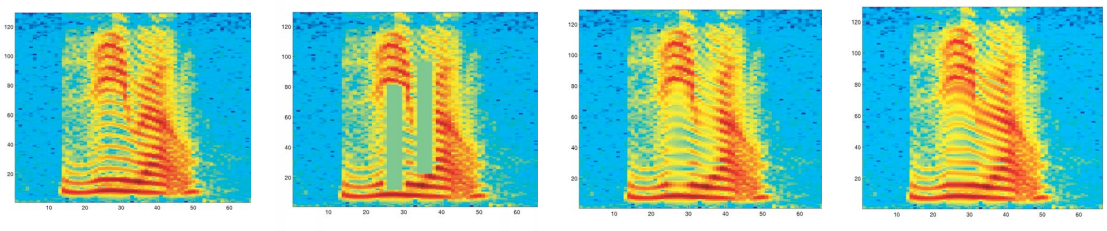


Figure 4: Illustration of the restoration (interpolation) of deleted time-frequency regions through the 'deformable spectrogram' model. The first pane shows the original signal (a single voice); in the second pane, two rectangular regions have been deleted. The final two panes show restoration early and late in the iterative inference procedure.

speech recognition model they are using). They find the solution for both states and segregation with greatest posterior probability by using a modified form of the Viterbi decoder from a conventional speech recognizer. This formulation includes a factor to reflect the likelihood of any given segmentation based directly on the observations (i.e. independent of the inferred states) – a natural place to link in the CASA features of section 2.2.

To be effective, speech signal models need to be tightly specific to the target speaker. In the prior work mentioned above, the speakers are known in advance, and the models have been based on a substantial amount of speaker-specific training data. Performance would be far worse for generic voices. Similar issues in speech recognition have been addressed by speaker adaptation, in which a generic model is 'sharpened' to match a particular speaker by estimating a few adaptation parameters. In order to be able to use detailed speech models for restoration of previously-unseen voices, we will adapt such an approach for the detailed spectral models used in Roweis [2003] etc.

Instead of attempting to memorize the possible realizations of source, a second approach is to build a model that encodes the possible dynamics of the signal independent of the instantaneous spectral detail. This model ought to be more compact, allowing more accurate inferences for a given model complexity and/or training set size. A recent instance of this approach is the 'deformable spectrogram' [Reyes-Gomez et al., 2004, 2005b], which models the evolution of spectral slices in a time-frequency surface as local transformations such as upward or downward translations. These transformations are modeled in a parallel layer of hidden states with slower variation (i.e. better predictability) than the spectra themselves. The dynamics of the transformation layer (modeled as a Markov random field with each node dependent on all its neighbors, and inferred with belief propagation) can be learned from generic speech material, then fit to a new speech observation. Then, any missing or corrupt observations in the original signal can be reconstructed first by interpolating the transformation layer across the missing regions, then using the transformations to 'grow back' the missing regions from the boundaries. Figure 4 illustrates the results of such an interpolation. Recent results have applied this technique to separating and reconstructing mixtures of sounds [Reyes-Gomez et al., 2005a]. Informally, this approach can generate restorations that sound very natural; in the context of this project, we will be able

to investigate the more difficult question of whether and how this kind of restoration can be used to improve intelligibility. We will also investigate combining this efficient description of dynamics with the rich inventory of speaker sounds from the memorization models.

We will make a careful study of the ways in which model-based separation fails: the kinds of conditions that lead to incorrect inference, and the kinds of mistakes that occur. Model-based inference can fail when the available observations result in too much ambiguity (for instance, when nothing is observed), but in these cases the posterior $p(\mathbf{x}_u)$ may still provide useful information for the processing in section 2.5. More problematic are situations in which incorrect inferences are made, for instance because noisy or unexpected observations match an inappropriate state: we will experiment to see when this happens, and how we can trade quality of correct inferences against risk of mistakes. Similarly, we will investigate the kinds of gross mistakes that signal interpolation schemes can make, for instance when the seemingly consistent edges of a missing region are connected as a reconstruction of a segment that was in fact much less uniform. We will examine empirically how often this occurs, and develop mechanisms to support the tradeoff between successful and incorrect interpolations.

2.4 WP4: Human Speech Intelligibility and Restoration

Perceptual studies will focus on evaluating the intelligibility of speech with the kinds of gaps and artifacts that come from separation algorithms, and how this can be affected if gaps are filled in by various types of noise, leading to the development of beneficial 'proxy noise' (noise which maximizes the likelihood that listeners can perceptually fill in the speech information blanked out in the gaps). Intelligibility is expected to depend on the amount of information provided by the speech context near the gaps (i.e., by word, phrase, sentence, and beyond) and also on the ability of the noise to stimulate the listener to recover the original information obliterated by the gap. The auditory-perceptual and linguistic-cognitive processes that allow the listener to fill in missing speech information are not well understood. A better understanding of the factors governing this human processing will both inform the development of signal processing methods that yield higher intelligibility and improve the perceptual models on which many CASA algorithms are based.

Our experiments will be conducted with young, normal-hearing college students with previous experience as subjects in psychoacoustic research. For each tested condition, we will present to each subject 60 sentences from the IEEE sentence lists [IEEE, 1969] and 100 digit strings selected from TIDIGITS [Leonard, 1984], and score percent correct sentences and words. We will also determine phonemic confusions, following in the footsteps of Miller and Nicely [1955], to identify systematic confusions between phonetic features that can provide specific feedback for algorithm improvement. Intelligibility of the reconstructed and filled-in speech will be compared to that for a baseline stimulus consisting of the same mixture of target and masking signals as was presented as input to the separation algorithm.

2.4.1 Proxy Noise

In a classic experiment, Miller and Licklider [1950] presented listeners with periodically interrupted speech. The interruptions degraded the quality of the speech, but adding noise to the gaps restored the subjective impression that the speech was continuous. Moreover, Bashford et al. [1992] and Warren et al. [1997] later demonstrated that this subjective impression was accompanied by an improvement in intelligibility. That noise can improve intelligibility seems a paradox, but it may be understood by noting that silent gaps provide strong evidence for an *absence* of sound. Filling the gaps weakens this misleading evidence. Furthermore, the gaps, or the transitions that bound them, may resemble phonemic cues that further mislead the listener; the added noise attenuates those cues.

In past experiments, the noise generally has a fixed spectral shape and level adjusted to be sufficient to promote restoration. However, it is likely that restoration can be improved by crafting the noise in a way that promotes restoration of the missing speech while minimizing energetic masking of surrounding target material. Most useful would be some kind of 'partially informative noise' that conveys the limited inferences concerning deletions with the minimum of information masking distraction. For example, suppose that models predict two equally plausible phonemes sharing one or more phonetic and/or suprasegmental features (such as vowel identity, pitch contour, etc.) in the missing slot. In such cases, it may be possible to design a hybrid sound that evokes either phoneme equally well while excluding any other. This hybrid approach would draw on the same modeling framework as the interpolation approach (section 2.3).

2.4.2 Specific aims

Efficacy of Machine Algorithms for Improving Speech Intelligibility Our first aim is to measure and compare intelligibility of reconstructed speech embedded in various types of maskers. In the baseline experiments, the same set of maskers, set at the same SNR, will be added to the target speech to create the input to the machine algorithm; therefore, intelligibility gain or loss caused by the algorithm processing can be determined for each masker condition and a selected set of gap fillers.

Effects of Gaps on Speech Intelligibility: Inserting silent gaps at random intervals will harm intelligibility, but such harm will probably be reduced if gaps are filled by noise and will vary with properties of the filler. We wish to measure systematically how statistics of missing speech and the density and nature of the gaps influence target speech intelligibility. In particular, we will compare gaps filled with silence, uninformative broad-band noise, partially informative noise sharing spectral, temporal, or pitch characteristics with the missing speech (as might be inferred from a model), and misinformative noise (as might result from feed-through from an interfering voice).

We predict that both the speech material characteristics (e.g. entropy) and the gap characteristics (shape, density and temporal distribution) will affect performance. Durations equal to or longer than a syllable should have the largest effect. We expect that *partially informative noise* is better than *uninformative noise*, itself better than *silence*, itself better

than *misinformative noise*. For partially informative noise (e.g. based on model inferences), we expect temporal structure to be more important than spectral structure, because the amplitude envelope of speech alone can yield highly intelligible speech [Shannon et al., 1995]. However, adding the pitch contour or the spectral shape of speech should produce even better performance. This hypothesis justifies efforts in WP2, WP3, and WP5, to recover such speech features even in cases where full segregation or reconstruction is impossible.

2.5 WP5: Integration, Presentation, and Evaluation

The goal of integration is to construct an output signal that will have the best intelligibility to a listener. There are a number of possible choices for this signal: the original mixture, the separated signals from the algorithms of WP2, the inferred restorations of corrupt regions from WP3, and the 'proxy noise' maskers arising from the studies in WP4. The decisions about which of these to use, and how to combine them if more than one signal is to be combined, will depend on the results of the 'failure mode' investigations: How well can we predict when particular algorithms give unfavorable or misleading outputs, and how severe will the impact of these failures be on listeners?

One choice is a simple selection scheme, which would involve developing signal dependent estimators of the intelligibility of the target speech for each of the different signal presentations, then dynamically choosing the signal with the best expected intelligibility. This, however, immediately raises the question of the granularity at which to make this decision: If the signal processing is switched within a single phrase or utterance, it is likely to introduce perceived artifacts that will distract listeners and degrade intelligibility; at the same time, we desire a system able to apply rapidly-varying, context-dependent processing of the signal, to take advantage of the locally most favorable conditions. For situations in which certain sounds are best processed one way, and others another, we will investigate the effects and trade-offs involved in different ways of combining them, for instance by cross-fades in time and frequency with different slopes.

A second way to combine the algorithms is in a single processing scheme that draws on all the components. Already, we have noted that the model-based inference of WP3 relies on an estimate of the observable ('known') signal components, and the local analysis of WP2 is an important source for this information. If CASA and multichannel processing techniques can enhance the target in some regions but identify other regions as beyond recovery, this is likely to be a better starting point for model-based interpolation than the original, unprocessed mixture. Knowing the points of uncertainty and potential error in the interpolations, as well as the location of residual interfering energy, will direct the placement of distraction-minimizing noise to defuse the disruption caused by these components.

The decision of whether to add noise (to promote perceptual restoration) or interpolate (to restore automatically) also depends on the ability or desire of the listener to perform restoration in difficult situations. If the listener's cognitive resources are limited – for example as a result of natural aging [Salthouse, 1991] – then interpolation may be preferable to reduce cognitive load on the listener. Devising methods to predict a listener's restoration abilities will be part of our psychoacoustic investigation.

The development of integration schemes will be directed by empirical results from psychacoustic testing, since we do not expect any simple signal properties will be adequate to predict the relative intelligibility of different processing options. Thus, this task will involve a close collaboration between the development of processing algorithms and psychoacoustic experimentation and evaluation. Through a combination of guided developments informed by diagnostic listener tests on the component separation candidates, and empirical tests to choose between specific design choices, the net result will be the best-performing intelligibility enhancement for target speech in the kinds of real-world multispeaker noise that are the focus of this project.

3 Education and Outreach

This proposal originated in a pair of NSF and AFOSR-sponsored workshops on speech separation organized by the proposers (for more details, see Divenyi's prior support in section 6). The rationale for the workshops was that researchers in many disciplines were considering the problem of how speech mixtures are separated or understood, and that they needed a forum to learn about one another's work. This project stems from the same motivation and is proposing a direct collaboration between the experimental psychology labs at Boston University and EBIRE, and the engineering/computer science groups at Columbia and Ohio State Universities (along with our consultant from CNRS in Paris who has a long track record of connecting signal processing techniques with perceptual results). A major benefit of the project will be the education of the graduate students at each institution who will naturally be drawn out of their disciplinary shells. Because we believe interdisciplinary experience is so important to making progress on the speech separation problem, we have included travel funds in our budget to enable a student exchange program: The graduate students supported on this project will each be obliged to spent at least one summer at one of the other partner labs (funding for travel to the Paris lab is being sought under a separate OISE proposal). In this way we hope to seed a new generation of researchers comfortable with both signal processing algorithms and experimental psychology protocols.

In addition to the directly-supported students, the project will impact the wider student bodies at the universities involved. PI Ellis is planning a new course based on the psychology and technology of speech separation as a spin-off from his successful Speech and Audio survey class [Ellis, 2005]; this will hopefully involve participation by specific interested colleagues in Columbia's Psychology and Biomedical Engineering departments. Ellis also teaches Digital Signal Processing to engineering undergraduates and graduates from several departments; audio demonstrations are a popular feature of this class, and speech separation will provide a stimulating source of examples. At BU, co-PI Shinn-Cunningham will develop a new module on speech restoration for inclusion in her course on hearing, including results and examples from both models and perceptual experiments. This course is cross-listed, serving as one of the required courses for graduate students in her home department of Cognitive and Neural Systems, and filling the undergraduate requirement for an advanced-level elective for Biomedical Engineering undergraduates. At EBIRE, co-PI Di-

venyi will continue to accept advanced engineering undergraduate students from across the US for summer internships in order to provide them with experience in a behavioral audio lab (last year's intern has been accepted in the Harvard-MIT Hearing and Speech program). Co-PI Wang recently introduced a new graduate course entitled 'Computational Audition' which covers computational models and algorithms for auditory perception. Progress in the proposed research will enhance this course and make it more appealing to speech/hearing and cognitive science students at OSU. Wang also plans to draw class project topics from several aspects of this research effort.

4 Work Plan

Year 1: WP1: Collection of field recordings, examination of statistics. WP2: Establishing baseline CASA/BSS algorithms and characterizing failure modes. WP3: Implement baseline state-based and dynamics models and characterize failure modes. WP4: Investigate effect of gaps on speech perception.

Year 2: WP1: Synthesize speech-in-noise material. WP2: Integrate local-feature separation algorithms. WP3: Speaker adaptation in predictive models. WP4: Develop and test 'proxy noise' fillers. WP5: Integrated separation and inference system.

Year 3: WP2, WP3: Characterize, refine, and integrate models. WP4: Experiments on influence of separation algorithm artifacts. WP5: Arbitration between different systems, tests of overall intelligibility improvement.

5 Impact and Conclusions

This proposal forms the basis of a critical larger project to address signal separation that both mimics human abilities, and delivers value to human users – something lacking in the majority of existing work. While the constraints of a 3 year project limit the scope of what we can tackle, we aim to effect a shift in the perspective of the entire community. In particular, listeners evidently employ many as-yet poorly understood principles to understand speech in noise, and further psychoacoustic experiments, guided and aided by computational models, will reveal these. Conversely, we do not expect that the integrated machine system we will produce will exhaust the potential of these algorithms, and the opportunities to combine the strengths of multiple approaches in different ways will likely open numerous research directions to extend well beyond the current project's lifetime.

A qualitative jump in the capabilities of speech separation and enhancement techniques will find applications in a huge range of activities, from enhancing old recordings for broadcast, through to restoring the full communicative ability of people with damaged hearing. By pursuing the most promising current approaches, at the same time as characterizing and working with their limitations, all oriented towards the objective of delivering a real improvement to listeners, the proposed project will take us most rapidly to a situation where speech separation can provide actual benefit to real users.

6 Results from prior support

PI **Ellis** is currently in the third year of NSF project IIS-0238301 "CAREER: The Listening Machine: Sound Source Organization for Multimedia Understanding" ($500,000, award period 2003-06-01 to 2008-05-31). This project is concerned with using machine learning to recognize different sources in sound mixtures, and led to the development of the "deformable spectrogram" model mentioned above [Reyes-Gomez et al., 2005b, 2004]. Current work includes automatic segmentation and classification of long-duration personal audio recordings into diary-style events [Ellis and Lee, 2004a,b].

Ellis was also a co-PI on NSF project IIS-0121396 "ITR/PE+SY: Mapping Meetings: Language Technology to make Sense of Human Interaction" ($1,402,851, award period: 2001-09-01 to 2005-08-31, PI: Nelson Morgan, ICSI). This project was concerned with the application of speech recognition and other signal analysis techniques to information extraction from natural meetings. Publications include work on 'episode' segmentation [Renals and Ellis, 2003], finding emphasized utterances and laugher [Kennedy and Ellis, 2003, 2004], and segmenting speaker turns from multi-microphone information [Ellis and Liu, 2004].

Co-Pi **Divenyi** was recipient of NSF Conference Award 0345301 "Perspectives on Speech Separation – A Workshop" for an invitational interdisciplinary workshop held between October 31 and November 2, 2003 in Montreal, Canada. At this workshop, 22 computer scientists, hearing and speech scientists, and neuroscientists presented their work and discussed issues and future plans regarding research on speech separation. A book containing the workshop presentations of the participants was published in late 2004 by Kluwer Academic Publishers [Divenyi, 2005]. An amended extension of this award was received from NSF to make it possible for eleven predoctoral and postdoctoral persons to attend and present their work at the AFOSR-sponsored workshop "Speech Separation and Comprehension in Complex Acoustic Environments by Humans and Machines", held between November 4 and 7, 2004, in Montreal, Canada.

Co-PI **Wang** was PI on NSF project IIS-0081058 "ITR/HCI: Dynamics-Based Speech Segregation" ($450,000, award period 2000-09-01 to 2004-08-31). This project has yielded major results, including a new robust algorithm for multipitch tracking of noisy speech by integrating periodicity information from automatically-chosen subsets of channels [Wu et al., 2003], a voiced speech segregation system that applies special treatment to high-frequency regions of unresolved harmonics, grouping them on the basis of amplitude modulation rates that match the estimated pitch of the target speech [Hu and Wang, 2004], and a model for separation based on spatial cues [Roman et al., 2003, 2004]; all models are comprehensively evaluated and shown to be substantially superior to earlier approaches. Two Ph.D. students supported on this project have successfully defended.

References

T. L. Arbogast, C. R. Mason, and G. Kidd, Jr. The effect of spatial separation on informational and energetic masking of speech. *J. Acoust. Soc. Am.*, 112:2086–2098, 2002.

F. R. Bach and M. I. Jordan. Blind one-microphone speech separation: A spectral learning approach. In *Advances in Neural Information Processing Systems (NIPS)*, volume 17. MIT Press, Cambridge MA, 2004. URL `http://cmm.ensmp.fr/~bach/fbach_nips_2004_speech.pdf`.

J. Barker, M. P. Cooke, and Daniel P. W. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45:5–25, 2005. URL `http://www.ee.columbia.edu/~dpwe/papers/BarkCE05-sfd.pdf`.

J. A. Bashford, K. R. Riener, and R. M. Warren. Increasing the intelligibility of speech through multiple phonemic restorations. *Percept. Psychophys.*, 51(3):211–217, 1992.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995. URL `citeseer.nj.nec.com/bell95informationmaximization.html`.

A. S. Bregman. *Auditory Scene Analysis*. Bradford Books, MIT Press, 1990.

G. J. Brown and D. L. Wang. Separation of speech by computational auditory scene analysis. In J. Benesty, S. Makino, and J. Chen, editors, *Speech Enhancement*. Springer, New York, 2005. in press.

D. S. Brungart and B. D. Simpson. Auditory localization of nearby sources in a virtual audio display. In *Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acoustics*, pages 107–110, Mohonk, 2001.

D. S. Brungart and B. D. Simpson. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. *J. Acoust. Soc. Am.*, 112: 664–676, 2002.

E. C. Cherry. Some experiments on the recognition of speech with one and two ears. *J. Acoust. Soc. Am.*, 25:975–979, 1953.

M. Cooke and D. P. W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35(3–4):141–177, 2001. `http://www.ee.columbia.edu/~dpwe/pubs/tcfkas.pdf`.

M. P. Cooke. 10+1 perspectives on speech separation and identification in listeners and machines, November 2004. URL `http://labrosa.ee.columbia.edu/Montreal2004/talks/cooke_casa.pdf`. Presentation slidepack.

M. P. Cooke. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, 2005. URL `http://www.dcs.shef.ac.uk/~martin/GlimpsingCookeJasa2005.pdf`. submitted.

M. P. Cooke, P. D. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, 2001. URL http://www.dcs.shef.ac.uk/research/groups/spandh/projects/respite/publications/cooke_speechcom_00.pdf.gz.

C. J. Darwin and R. W. Hukin. Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *J. Acoust. Soc. Am.*, 102:2316–2324, 1997.

C. J. Darwin and R. W. Hukin. Auditory objects of attention: The role of interaural time diffferences. *Journal of Experimental Psychology: Human Perception and Performance*, 25:617–629, 1999.

C. J. Darwin and R. W. Hukin. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J. Acoust. Soc. Am.*, 107:970–977, 2000.

A. de Cheveigné. The cancellation principle in acoustic scene analysis. In P. Divenyi, editor, *Speech separation by humans and machines*, pages 245–259. Kluwer, Norwell MA, 2004.

A. de Cheveigné and A. Baskind. F0 estimation of one or several voices. In *Proc. Eurospeech-03*, pages 833–836, Geneva, 2003.

Digital Innovations. The Neuros digital audio computer, 2003. http://www.neurosaudio.com/.

P. L. Divenyi, editor. *Speech Separation by Humans and Machines*. Kluwer Academic Publishers, New York, 2005.

N. I. Durlach, C. R. Mason, G. Kidd, Jr., T. L. Arbogast, H. S. Colburn, and B. G. Shinn-Cunningham. Note on informational masking. *J. Acoust. Soc. Am.*, 113:2984–2987, 2003.

D. P. W. Ellis. Evaluating speech separation systems. In P. Divenyi, editor, *Speech Separation by Humans and Machines*, chapter 20, pages 295–304. Kluwer, 2004. URL http://www.ee.columbia.edu/~dpwe/pubs/Ellis04-sepeval.pdf.

D. P. W. Ellis. ELEN E6820: Speech and Audio Processing and Recognition, 2005. URL http://www.ee.columbia.edu/~dpwe/e6820/. Online course materials.

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 2004a. http://www.ee.columbia.edu/~dpwe/pubs/sapa04-persaud.pdf.

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 2004b. URL http://www.ee.columbia.edu/~dpwe/pubs/carpe04-minimpact.pdf.

D. P. W. Ellis and J. Liu. Speaker turn segmentation based on between-channel differences. In *Proc. NIST Meeting Recognition Workshop*, Montreal, March 2004. URL `http://www.ee.columbia.edu/~dpwe/pubs/nist04-turnid.pdf`.

N. French and J. Steinberg. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.*, 19(1):90–119, 1947.

W. G. Gardner. Reverberation algorithms. In M. Kahrs and K. Brandenburg, editors, *Applications of Digital Signal Processing to Audio and Acoustics*. Kluwer Academic Publishers, Norwell, MA, 1998.

L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, AP-30(1), January 1982.

M. L. Hawley, R. Y. Litovsky, and J. F. Culling. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *J. Acoust. Soc. Am.*, 115:833–843, 2004.

G. Hu and D. L. Wang. Separation of stop consonants. In *Proc. IEEE Int. Conf. Acous., Speech,, Sig. Processing*, pages II.749–752, Hong Kong, 2003.

G. Hu and D. L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, 15:1135–1150, 2004.

G. Hu and D. L. Wang. Separation of fricatives and affricates. In *Proc. IEEE Int. Conf. Acous., Speech,, Sig. Processing*, pages II–749–752, Philadelphia, 2005.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Wiley, New York, 2001.

IEEE. *IEEE recommended practice for speech quality measures*. Institute of Electrical and Electronic Engineers, New York, 1969.

L. Kennedy and D. P. W. Ellis. Pitch-based emphasis detection for characterization of meeting recordings. In *Proc. Automatic Speech Recognition and Understanding Workhop IEEE ASRU 2003*, December 2003.

L. Kennedy and D. P. W. Ellis. Laughter detection in meetings. In *Proc. NIST Meeting Recognition Workshop*, Montreal, March 2004. URL `http://www.ee.columbia.edu/~dpwe/pubs/nist04-laughs.pdf`.

G. Kidd, Jr., C. R. Mason, and A. Brughera. The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acta Acustica*, 2005. in press.

B. Kollmeier and R. Koch. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *J. Acoust. Soc. Am.*, 95 (3):1593–1602, 1994.

H. Krim and M. Viberg. Two decades of array signal processing research: The parametric approach. *IEEE Signal Processing Magazine*, 13:67–94, 1996.

T. W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 6:87–90, 1999.

R. G. Leonard. A database for speaker independent digit recognition. In *Proc. IEEE Int. Conf. Acous., Speech,, Sig. Processing*, pages III–42–45, 1984.

C. Liu, B. C. Wheeler, Jr. W. D. O'Brien, C. R. Lansing, R. C. Bilger, D. L. Jones, and A. S. Feng. A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interfaces. *J. Acoust. Soc. Am.*, 110:3218–3231, 2001. URL `http://www.brl.uiuc.edu/Publications/2001/Liu-JASA-32182001.pdf`.

R. Meddis. Simulation of auditory-neural transduction: Further studies. *J. Acoust. Soc. Am.*, 83:1056–1063, 1988.

G. A. Miller and J. C. R. Licklider. The intelligibility of interrupted speech. *J. Acoust. Soc. Am.*, 22:167–173, 1950.

G. A. Miller and P. E. Nicely. An anlysis of perceptual confusions among some english consonants. *J. Acoust. Soc. Am.*, 27:338–352, 1955.

N. Morgan, D. Baron, J. Edwards, Daniel P. W. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at ICSI. In *Proc. Human Lang. Techn. Conf.*, pages 246–252, 2001. URL `ftp://ftp.icsi.berkeley.edu/pub/speech/papers/hlt01-mr.pdf`.

T. Nakatani and M. Miyoshi. Blind dereverberation of single channel speech signal based on harmonic structure. In *Proc. IEEE Int. Conf. Acous., Speech,, Sig. Processing*, pages I–92–95, 2003.

K. J. Palomäki, G. J. Brown, and D. L. Wang. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication*, 43:361–378, 2004.

R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice. Svos final report, part b: Implementing a gammatone filterbank. Technical Report 2341, MRC Applied Psychology Unit, 1988.

S. Renals and D. P. W. Ellis. Audio information access from meeting rooms. In *Proc. IEEE Int. Conf. Acous., Speech,, Sig. Processing*, 2003. URL `http://www.dcs.shef.ac.uk/~sjr/pubs/2003/icassp03-mtg.html`.

M. Reyes-Gomez, D. P. W. Ellis, and N. Jojic. Signal separation and restoration with deformable spectrograms, 2005a. in preparation.

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. URL `http://www.ee.columbia.edu/~dpwe/pubs/sapa04-transform.pdf`.

M. Reyes-Gomez, N. Jojic, and D. P. W. Ellis. Deformable spectrograms. In *Proc. AI and Statistics*, Barbados, 2005b. URL `http://www.ee.columbia.edu/~dpwe/pubs/aistats05-defspec.pdf`.

S. Rickard and O. Yilmaz. On the approximate W-disjoint orthogonality of speech. In *Proc. IEEE Int. Conf. Acous., Speech,, Sig. Processing*, volume 1, pages 529–532, Orlando FL, 2002.

N. Roman, D. L. Wang, and G. Brown. Speech segregation based on sound localization. *J. Acoust. Soc. Am.*, 114:2236–2252, 2003.

N. Roman, D. L. Wang, and G. Brown. A classification-based cocktail-party processor. In *Advances in Neural Information Processing Systems (NIPS)*, volume 14. MIT Press, Cambridge MA, 2004.

S. Roweis. One-microphone source separation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 11, pages 609–616. MIT Press, Cambridge MA, 2001.

S. Roweis. Factorial models and refiltering for speech separation and denoising. In *Proc. EuroSpeech*, Geneva, 2003.

T. A. Salthouse. *Theoretical perspectives on cognitive aging*. Lawrence Erlbaum, Hillsdale NJ, 1991.

R. V. Shannon, F.-G. Zeng, J. Wygonski, V. Kamath, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270:303–304, 1995. URL `http://www.ucihs.uci.edu/hesp/publications/zeng13.pdf`.

B. G. Shinn-Cunningham. Speech intelligibility, spatial unmasking, and realism in reverberant spatial auditory displays. In *Proc. Int. Conf. on Auditory Display*, pages 419–423, Kyoto, 2002.

B. G. Shinn-Cunningham, A. Ihlefeld, and M. Schoolmaster. Selective and divided attention: Extracting information from simultaneous sound sources. In *Proc. Int. Conf. on Auditory Display*, Sydney, Australia, 2004. URL `http://www.icad.org/websiteV2.0/Conferences/ICAD2004/papers/shinn-cunningham_etal.pdf`.

M. Slaney and R. F. Lyon. On the importance of time – A temporal representation of sound. In M. Cooke, S. Beet, and M. Crawford, editors, *Visual Representations of Speech Signals*. John Wiley, 1992.

B. D. van Veen and K. M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, April 1988.

R. M. Warren. Perceptual restoration of missing speech sounds. *Science*, 167:392–393, 1970.

R. M. Warren, K. R. Hainsworth, B. S. Brubaker, Jr. J. A. Bashford, and E. W. Healy. Spectral restoration of speech: intelligibility is increased by inserting noise in spectral gaps. *Perception and Psychophysics*, 59:275–283, 1997.

S. Winter, H. Sawada, S. Araki, and S. Makino. Overcomplete bss for convolutive mixtures based on hierarchical clustering. In *Proc. ICA*, pages 652–660, 2004.

M. Wu and D. L. Wang. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 2005. in press.

M. Wu, D. L. Wang, and G. J. Brown. A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, 11:229–241, 2003.