MAS.632 Conversational Computer Systems
Fall 2008

A Review of The Cocktail Party Effect

Barry Arons
MIT Media Lab
20 Ames Street, E15-353
Cambridge MA 02139
barons@media-lab.mit.edu

Abstract

The "cocktail party effect"—the ability to focus one's listening attention on a single talker among a cacophony of conversations and background noise—has been recognized for some time. This specialized listening ability may be because of characteristics of the human speech production system, the auditory system, or high-level perceptual and language processing. This paper investigates the literature on what is known about the effect, from the original technical descriptions through current research in the areas of auditory streams and spatial display systems.

The underlying goal of the paper is to analyze the components of this effect to uncover relevant attributes of the speech production and perception chain that could be exploited in future speech communication systems. The motivation is to build a system that can simultaneously present multiple streams of speech information such that a user can focus on one stream, yet easily shift attention to the others. A set of speech applications and user interfaces that take advantage of the ability to computationally simulate the cocktail party effect are also considered.

Introduction

> *"One of the most striking facts about our ears*
> *is that we have two of them—*
> *and yet we hear one acoustic world;*
> *only one voice per speaker"* [CT54]

This paper investigates aspects of *selective attention* in the auditory system—under what conditions can a listener attend to one of several competing messages? Humans are adept at listening to one voice in the midst of other conversations and noise, but not all the mechanisms for this process are completely understood. This attentional ability has been colloquially termed the *cocktail party effect* [Han89].

The phenomenon can be viewed in many ways. From a listener's point of view, the task is intuitive and simple. From a psychological or physiological perspective there is a vast and complex array of evidence that has been pieced together to explain the effect—there are many interactions between the signal, the auditory system, and the central nervous system. Acoustically, the problem is akin to separating out a single talker's speech from a spectrogram containing signals from several speakers under noisy conditions. Even an expert spectrogram reader would find this task impossible [Bre90].

Most of the evidence presented has been obtained from perceptual experiments that have been

performed over the last 40-odd years. Unfortunately, such perceptual evidence is often not as quantifiable as, for example, physical resonances of the vocal tract. Therefore, the bulk of the ideas and experimental results presented are qualitative, and an "exact" solution to the cocktail party problem cannot be found. While the focus of the paper is on voice signals and speech communication, note that much of the low-level perceptual evidence is based on experiments using simple stimuli, such as clicks, pure tones, or noise.

The Separation of Speech Channels

The cocktail party effect can be analyzed as two related, but different, problems. The primary problem of interest has traditionally been that of *recognition*: how do humans segregate speech sounds, and is it possible build a machine to do the task. What cues in the signal are important for separating one voice from other conversations and background noise? Can, and should, a machine use the same cues for the task, or can it use other acoustical evidence that humans are not efficient at detecting?

The inverse problem is the *synthesis* of cues that can be used to enhance a listener's ability to separate one voice from another in an interactive speech system. In a user interface it may be desirable to present multiple digitized speech recordings simultaneously, providing browsing capabilities while circumventing the time bottleneck inherent in speech communication because of the serial nature of audio [Aro91, SA89]. Synthesis of perceptual cues by a machine for human listeners might allow an application to perceptually nudge the user, making it easier to attend to a particular voice, or suggest that a new voice come into focus.

Early Work

Much of the early work in this area can be traced to problems faced by air traffic controllers in the early 1950's. At that time, controllers received messages from pilots over loudspeakers in the control tower. Hearing the intermixed voices of many pilots over a single loudspeaker made the controller's task very difficult [KS83].

Recognition of Speech With One and Two Ears

In 1953, Cherry reported on objective experiments performed at MIT on the recognition of messages received by one and two ears [Che53]. This appears to be the first technical work that directly addresses what the author termed the "cocktail party problem." Cherry proposed a few factors that may ease the task of designing a "filter" that could separate voices:

1. The voices come from different directions
2. Lip-reading, gestures, and the like
3. Different speaking voices, mean pitches, mean speeds, male vs. female, and so forth
4. Different accents
5. Transition probabilities (based on subject matter, voice dynamics, syntax . . .)

All factors, except for the last, can be removed by recording two messages from the same talker on magnetic tape. The author stated "the result is a babel, but nevertheless the messages may be separated." In a Shannonesque analysis, Cherry suggested that humans have a vast memory of transition probabilities that make it easy for us to predict word sequences [SW63].

A series of experiments were performed that involved the "shadowing" of recordings; the subject repeated words after hearing them from a tape recording. The contents of the recordings were often related, and in the same style, such as by selecting adjacent paragraphs from the same book. Recognition was often in phrases, and the subjects found the task very difficult, even though the recordings could be repeated an unlimited number of times. In no cases were any long phrases (more than 2–3 words) incorrectly identified, and the errors made were typically syntactically correct. In a slight variant of the setup, the subject was allowed to make notes with a pencil and paper. This long-term memory aid made the task much easier, and time required to perform the task was shortened—the messages were almost entirely separated by the subject.

In a similar experiment, the spoken phrases were composed of strings of clichés strung together with simple conjunctions and pronouns[1]. These artificially constructed "highly probable phrases" were nearly impossible to separate. Because the transition probabilities between phrases were low, the subject would select phrases equally from the two speech streams.

Subjects also listened to different spoken messages presented to each ear with headphones. In this configuration there is no directionality, there is simply a dichotic signal. The subjects had no difficulty in listening to the message played to one ear while rejecting sounds in the other ear. The recognition process can easily be switched to either ear at will. The subject could readily shadow one message while listening, though with a slight delay. Norman states that "the longer the lag, the greater advantage that can be taken of the structure of the language" [Nor76]. Note that the subject's voice is usually monotonic and they typically have little idea of the content of the message in the *attended* to ear. Virtually nothing can be recalled about the message content presented to the other (rejected) ear, except that sounds were occurring.

> This is what might be called the "what-did-you-say" phenomenon. Often when someone to whom you were not "listening" asks you a question, your first reaction is to say, "uh, what did you say?" But then, before the question is repeated, you can dredge it up yourself from memory. When this experiment was actually tried in my laboratory, the results agreed with our intuitions: there is a temporary memory for items to which we are not attending, but as Cherry, James, and Moray point out, no long-term memory. ([Nor76] page 22)

In follow-up experiments, the language of the signal in the rejected ear was switched to German (by an English speaker), but the subjects did not notice the change. Changes from male to female speaker were usually identified, and a change to a pure tone was always identified. Reversed speech, such as a tape played backwards (having the same spectrum as the original signal, but no semantic content), was identified as having "something queer about it" by a few listeners, but was thought to be normal speech by others. In summary, the broad statistical properties of the signal in the rejected ear were recognized, but details such as language, individual words, and semantic content were unnoticed.

[1] The texts were generated from 150 clichés from speeches reported in the newspapers. For example: "I am happy to be here today to talk to the man in the street. Gentlemen, the time has come to stop beating around the bush—we are on the brink of ruin, and the welfare of the workers and of the great majority of the people is imperiled."

In an interesting variant of these studies, the same recording was played to both ears with a variable delay between the ears. The experiment would proceed as above, with the subject shadowing one recording. The time delay was slowly decreased, until at a point when the recordings were within 2–6 seconds of each other, the subject would exclaim something like "my other ear is getting the same thing." Nearly all the subjects reported that at some point they had recognized that words or phrases in the rejected ear were the same as in the attended ear. Note that this result is surprising in light of the previous tests where the subjects were unable to identify even a single word in the rejected ear.

By switching one message periodically between the ears, the time interval needed to transfer attention between the ears was determined. For most subjects this interval was about 170 ms. A further study investigates this in more detail, defining τ as the average "word recognition delay" [CT54]. Note that τ represents the *entire* complex hearing process, and is not just because of the sensory system.

Responding to One of Two Simultaneous Messages

Spieth *et al.* at the Navy Electronics Laboratory in San Diego performed a series of experiments investigating responses to the presentation of simultaneous messages [SCW54]. The goal of the first set of experiments was to find conditions under which a communication's operator could best recognize and attend to one speech message when it was presented simultaneously with another irrelevant message. Communication messages do not provide visual cues to aid in the identification of the sender or the perception of the message. While redundancy within a message is high, competing messages are of similar form, content, and vocabulary.

Several configurations were tried that presented messages with horizontally separated loudspeakers. It was found that three loudspeakers (at $-10°$, $0°$, and $+10°$ azimuth) increased channel identification scores over a single loudspeaker (at $0°$ azimuth), and that a larger separation ($-90°$, $0°$, and $+90°$ azimuth) improved scores further[2]. Variants of this experiment were performed (e.g., with added visual cues, low-pass filtering the messages, etc.), and an increased horizontal separation always reliably improved scores.

Messages that were high- and low-pass filtered at 1.6 kHz, improved the operator's ability to answer the correct message and identify the channel. Note that the filtering did not significantly decrease the intelligibility of the messages. Both the high- or low-pass messages were made easier to attend to, and they could be separated from an unfiltered message.

Spieth relates this phenomenon to Cherry's work on transition probabilities: "this suggests the possibility that anything which increases the element-to-element predictability within each of two competing messages and/or decreases the predictability from an element in one stream to a succeeding element in the other stream, will make either stream easier to listen to." Note that this fundamental theme resurfaces throughout many of the studies. The authors propose that further narrowing the frequency bands, and increasing the separation between them will further improve the ability to listen to either stream. This is, however, limited by the point at which the bandwidth is so narrow, or frequency so extreme, that intelligibility of the individual messages is impaired.

If two or more separation aids were used at the same time (e.g., filtering *and* spatial separation),

[2]Correct identification scores for a particular task under these three conditions increased from 76% to 86% to 96%.

scores were usually improved with respect to a single aid, but the effect was not fully additive. The authors hypothesize that the reason the effects were not additive was because of the general ease of the tasks (i.e., it was not difficult to achieve a score of 100%).

Responding to Both of Two Simultaneous Messages

A related study by Webster and Thomas investigated responding to *both* of two overlapping messages [WT54]. As in the previous experiment, more correct identifications for *sequential* messages were found using six loudspeakers than one. Having a "pulldown" facility (the ability to manually switch the audio from one particular loudspeaker to a headphone or near-field loudspeaker) gave considerably better results. It was also found that the louder of the two simultaneous messages was more likely to be heard correctly. Note, however, that having multiple loudspeakers did *not* improve results when it was necessary to attend to two competing *simultaneous* messages.

The ability to rapidly shift one's attention (e.g., with multiple loudspeakers) does not help if the information rate is high. Under the worst conditions (two simultaneous messages), only 60% of the information was received, but this results in a greater total information intake per unit time than if the messages had occurred sequentially.

Selective Listening to Speech

In 1958, Broadbent summarized much of this early work, including his own experiments, and that of a variety of other researchers [Bro58]. It had been experimentally established by that time that the probability of a listener correctly hearing a word varies with the probability of the word occurring in a particular context. For example, after hearing the word "bread", the subsequent occurrence of "butter" or "knife" is more likely than "eraser" or "carburetor". In 1951 it was shown that a word is less likely to be heard correctly if the listener knew that it was one of many alternatives as compared with a small number. The performance of selective listeners thus seems to vary with *information as defined by communication theory*, rather than with the amount of physical stimulation.

Broadbent concludes from Webster's experiments that messages containing little information can be dealt with simultaneously, while those with high information content may not. He notes that the statement "one cannot do two tasks at once" depends on what is meant by "task." It is pointed out that spatial separation is helpful in situations that are similar to the task of the listener ignoring one channel and responding to the other—the spatial effect is less important when the listener is dealing with two channels simultaneously. Note also that the time to shift attention is *increased* when two messages come from different directions, and that this may cancel out other advantages of spatial separation.

Broadbent summarizes the three main conclusions of the selective listening experiments as:

1. Some central nervous system factors, rather than sensory factors are involved in message selection.
2. Effects vary with information content of the messages.

3. When information must be discarded, it is not discarded at random. If some of the information is irrelevant, it is better for it to come from a different place, to be at a different loudness, to have different frequency characteristics, or to be presented to the eye instead of the ear. When no material is to be discarded, there is little advantage in using two or more sensory channels for presenting information.

Binaural Unmasking

Our ability to detect a signal in a background masking signal is greatly improved with two ears. Under ideal conditions, the detection threshold for binaural listening will exceed monaural listening by 25 dB [DC78]. Consider, for example, a control condition where a signal and noise are played to a single ear. If the signal is then played simultaneously to both ears, but the phase of the noise to one ear is shifted by 180° with respect to the other ear, there is a 6 dB improvement in the detectability of the signal. This improvement over the control condition is called the binaural masking level difference (BMLD or MLD). If the noise is played to both ears, but the signal to the ears is 180° out of phase, there is a 15 dB BMLD.

The cocktail party effect can thus be partly explained by BMLD's. When listening binaurally, the desired signal coming from one direction is less effectively masked by noise that originates in a different direction [Bla83]. Such a technique is often exploited in earphones for fighter pilots to help separate speech signals from the high noise level of the cockpit. The headphones are simply wired so that the signal presented to one ear is antiphasic (180° out of phase) with the signal presented to the other ear.

Auditory Scene Analysis

A great variety of research relating to perceptual grouping of auditory stimuli into streams has recently been performed, and summarized, by Bregman [Bre90]. In the introduction to his book, Bregman talks about perceptual constancies in audition, and how they relate to vision:

> A friend's voice has the same perceived timbre in a quiet room as at a cocktail party. Yet at the party, the set of frequency components arising from that voice is mixed at the listener's ear with frequency components from other sources. The total spectrum of energy that reaches the ear may be significantly different in different environments. To recognize the unique timbre of the voice we have to isolate the frequency components that are responsible for it from others that are present at the same time. A wrong choice of frequency components would change the perceived timbre of the voice. The fact that we can usually recognize the timbre implies that we regularly choose the right components in different contexts. Just as for visual constancies, timbre constancy will have to be explained in terms of a complicated analysis by the brain, and not merely in terms of a simple registration of input by the brain.
>
> There are some practical reasons for trying to understand this constancy. There are engineers that are currently trying to design computers that can understand what a person is saying. However in a noisy environment, the speaker's voice comes mixed with other sounds. To a naive computer, each different sound that the voice comes mixed with makes it sound as if different words were being spoken, or as if they were spoken by a different person. The machine cannot correct for the particular listening conditions as the human can. If the study of human audition were able to lay bare the principles that govern the human skill, there is some hope that a computer could be designed to mimic it. ([Bre90] page 2)

Scene analysis in audition is concerned with the perceptual questions of deciding how many sound sources there are, what are the characteristics of each source, and where each source is located [Han89]. A baby, for example, imitates its mother's voice, but does not insert the cradle squeaks that have occurred simultaneously with the mother's speech. The baby rejects the squeaks as not being part of the perceptual object formed by the mother's voice—the infant has solved the scene analysis problem in audition. Bregman also states the problem a different way: ". . . it would be convenient to be able to hand a spectrogram over to a machine that did the equivalent of taking a set of crayons and coloring in, with the same color, all the regions on the spectrogram that came from the same source." This is what auditory scene analysis is all about.

Sounds or *acoustic events* are created when physical things happen. The *perceptual unit* that represents such a single happening is called an *auditory stream*. A series of footsteps, for example, each represent individual sounds, yet are usually experienced as a single perceptual event. Streams are a way of putting sensory information together. If the properties "far" and "lion roar" are assigned to one auditory stream, and "near" and "crackling fire" assigned to a different stream, we will probably behave differently than if the distance percepts were reversed [Bre90, Han89].

Many of the ideas of auditory scene analysis can be traced back to visual work done by the Gestaltists of the early 1900's [Han89]. Visual and auditory events are combined to make the most coherent perceptual objects. Elements belonging to one stream are maximally similar and predictable, while elements belonging to different streams are maximally dissimilar. The Gestalt psychologists organizational principles of the *visual* field include:

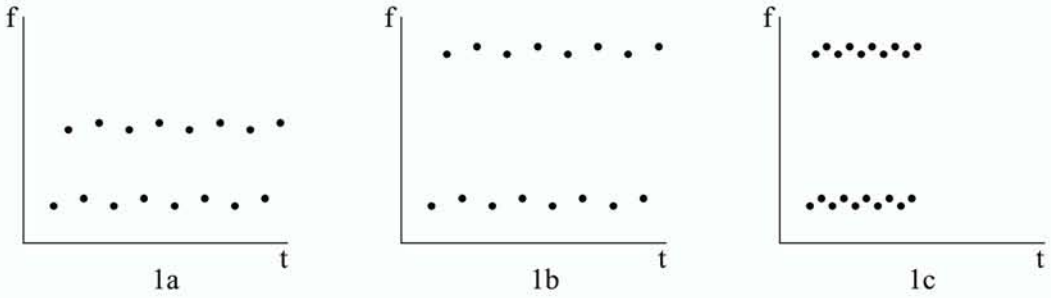
Similarity: elements that are similar in physical attributes tend to be grouped
Proximity: elements that are close together in space or time tend to be grouped
Continuity: elements that appear to follow in the same direction tend to be grouped
Common Fate: elements that appear to move together tend to be grouped
Symmetry & Closure: elements that form symmetrical and enclosed objects tend to be grouped

From this perspective, we expect acoustic events that are grouped into one perceptual stream to be similar (e.g., in frequency, timbre, intensity), to be in spatial or temporal proximity, and to follow the same temporal trajectory in terms of frequency, intensity, position, rhythm, etc.

Primitive Segregation

The focus of Bregman's work is on primitive, or unlearned, stream segregation. The following sections qualitatively summarize many of Bregman's findings that are relevant to the cocktail party effect. These ideas begin with general attributes of auditory scene analysis, and will move toward, and emphasize, the perception of speech streams.

Grouping Processes. There are two classes of grouping processes that can be broadly classified as *simultaneous integration* and *sequential integration* (these can also be called *spectral grouping* and *temporal grouping*). The following figures visually illustrate these types of groupings (the circles represent sounds at a particular frequency). In the figure below (after [Bre90]), the segregation is stronger in figure 1b than figure 1a, as the frequency separation between the high and low tones is greater. Similarly, the segregation is still greater in figure 1c where there is an increase in speed. The tones are more tightly packed in both the visual representation and the auditory stimuli.



Spatial Location. Primitive scene analysis groups sounds coming from the same location and segregates sounds that originate in different locations. As Cherry and others showed, a person can do a good job of segregating sounds from monaural recordings. Spatial cues are strongest when they are combined with other auditory cues—spatial evidence is just one cue in the complex scene analysis system. Note also that reflections (e.g., room, body) can significantly alter received acoustical signals.

> Engineers working in the automatic segregation of concurrent sounds have used spatial separation as a uniquely powerful way of determining whether the sounds have come from the same physical event (usually a talker). Humans use spatial origin too, but do not assign such an overwhelming role to it. They can do quite well at segregating more than one stream of sound coming from a single point in space, for example, from a single loudspeaker. ([Bre90] page 644)

Spatial Continuity. Sound sources (talkers) and listeners don't move too far or too fast. Experiments have shown that spatial discontinuities break down streams, so spatial continuities must be important at holding streams together.

Loudness differences. Differences in loudness may not, in themselves, cause segregation, but as with spatial location, such cues may strengthen other stream segregation evidence.

Continuity. Sounds hold together in a single stream better than discontinuous sounds. This continuity can be in fundamental frequency, temporal proximity, shape of spectra, intensity, or spatial origin. It is unlikely that a sound will begin at the same instant that another sound ends—when the spectra of incoming sensory data change suddenly, we conclude that only one sound has started or stopped.

A complicated spectrum, for example, may have a simpler spectrum embedded in it that was heard earlier. This simpler spectrum may be adjacent to the more complicated spectrum with no discontinuity. It is therefore reasonable to consider the part of the spectrum that matches the earlier one as a continuation of it, and treat the latter portion as resulting from the addition of a new sound to the mixture.

Visual Channel Effects. We tend to perceive sounds as coming from locations of visual events. Think of the illusion when watching television or a movie, where an actor's voice appears to be emanating from his mouth regardless of where the loudspeaker is located.

> An example of the interrelationship is that the grouping of sounds can influence the grouping of visual events with which they are synchronized and vice versa. . . the tendency to

> experience a sound as coming from a location at which visual events are occurring at the same temporal pattern (the so-called ventriloquism effect) can be interpreted as a way in which visual evidence about the location of an event can supplement unclear auditory evidence. The direction of influence is not just from vision to audition, but in the reverse direction as well. ([Bre90] page 653)

> Thus our interpretation of auditory spatial cues is strongly influenced by our perceived visual orientation. Or, more correctly, the highest level of spatial representation involves an integration of information from the different senses. ([Moo89] page 224)

History. Stream analysis processes use history to adjust for momentary spatial estimates. We use the fact that sounds and objects tend to move slowly in space and time and hence cause coherent structure.

Segregation Time Constant. It takes at least four seconds to build up and segregate a stream, and four seconds for it to go away after the sequence stops. This long time constant probably prevents the auditory system from oscillating under ambiguous conditions. However, a sudden change in the properties of a signal can reset the streaming mechanism more quickly than can silence.

Harmonics and Frequency Modulation. The perceived pitch of a complex tone depends on an estimate of the fundamental frequency of the set of harmonics that make up the tone (even if the fundamental is missing). The scene analysis mechanisms favor the grouping of harmonics of the same fundamental. Thus if several fundamentals account for all harmonics, we conclude that there are several sound sources.

> When the pitch rises, not only does the fundamental frequency go up but all the harmonics go up by the same proportion too. It is plausible to believe that this correlated change, if it could be detected auditorily, could tell us that the changing partials all came from the same voice. The auditory system could group all such correlated changes and hear only one changing sound.
>
> There is evidence to suggest that two types of frequency change (or modulation) are used for this purpose. One is micromodulation, the tiny fluctuations of the pitch of human voices that occur even when the speakers think they are holding a steady pitch . . . The other type of frequency modulation is the slow kind that occurs when we voluntarily vary the pitch of our voice in a smooth way as we do, for example, when we raise out pitch at the end of a question . . . The synchronization of the micromodulation or of slow modulation in different parts of the spectrum seems to cause those parts to be treated as parts of a single sound. ([Bre90] page 657)

Weighting of Evidence. There is collaboration, as well as competition, among the features used in a stream segregation decision. If the number of factors that favor a particular grouping of sounds is large, the grouping will be strong, and all the sounds will be heard as part of the same stream.

Schema-based segregation

Segregation that is learned, or involves attention, is considered to be based on a higher level of central processing. Anything that is consciously "listened for" is part of a *schema*. Recall from the findings of the earlier studies, that only a limited number of things can be attended to simultaneously, so there is a limitation on our ability to process schemas.

Primitive segregation is symmetrical. When it separates sounds by frequency (or location), we can attend to either high tones or low tones (left or right) equally well. Schema-based recognition is not symmetrical. If your name is mixed with other sounds it may be easy to recognize it in the mixture, but it does not make it easier to identify the other elements of the sound.

An example of the use of schema-based reasoning involves the simultaneous presentation of two synthetic vowels. The vowels were produced such that they had the same fundamental, the same start and stop time, and came from the same spatial location. All the primitive preattentive clustering theories suggest that these complex sounds should be fused into a single stream. However, higher level schema are used to distinguish the vowels in this mixture. Bregman suspects that the schema for each vowel is picking out what it needs from the total spectrum rather than requiring that a partitioning be done by the primitive processes.

There is also evidence that a scene that has been segregated by primitive processes can be regrouped by schemas. For example, a two-formant speech sound was synthesized with each formant constructed from harmonics related to a different fundamental. Listeners will hear two sounds, one corresponding to each related group of harmonics, yet at the same time, they will perceive a single speech sound formed by the complete set of harmonics. The speech recognition schemas thus can sometimes combine evidence that has been segregated by the primitive process.

Speech Scene Analysis

In addition to the grouping processes already mentioned, there are additional extensions and ideas that are specific to the analysis of speech signals. Note that it is often difficult to separate primitive processes from schema, and that speech schemas tend to obscure the contributions of primitive processes.

Considering the primitive segregation rules, it is somewhat surprising that voices hold together at all. Speech consists of sequences of low frequency complex tones (vowels) intermixed with high frequency noise (fricatives). With a production rate of roughly 10 phonemes/sec, speech should break up into two streams of alternating high and low tones. Listeners are able to understand and repeat a rapid sequence of speech, but are not able to report the order of short unrelated sounds (e.g., a hiss, buzz, etc.) played in sequence, even if they are played at a much slower rate than the corresponding phonemes.

> Warren argues that listeners to a cycle of unrelated events have to decompose the signal into constituent parts, recognize each part, and then construct a mental representation of the sequence. Listeners to speech do not have to go through this process ~~-th~~ey can do some global analysis of the speech event and match it to a stored representation of the holistic pattern. After all, Warren continues, children can recognize a word and often have no idea of how to break it up into its constituent phonemes. ([Bre90] page 534)

Pitch Trajectory. In general, the pitch of a speaker's voice changes slowly, and it follows melodies that are part of the grammar and meaning of a particular language. Listeners use both constraints to follow a voice over time.

In shadowing experiments two interesting results were shown. First, if the target sound and the rejected sound suddenly switched ears, the subjects could not prevent their attention from

following the passage (rather than the ear) that they were shadowing. The author of the original research argued that "the tracking of voices in mixtures could be governed by the meaning content of the message." Secondly, if only the pitch contour was switched between ears, subjects often repeated words from the rejected ear, even if the semantic content did not follow. The continuity of the pitch contour was, to some degree, controlling the subject's attention.

Spectral Continuity. Since the vocal tract does not instantaneously move from one articulatory position to another, the formants of successive sounds tend to be continuous. These coarticulatory features provide spectral continuity within and between utterances. Continuities of the fundamental and the formant frequencies are important at keeping the speech signals integrated into a single stream.

Pitch-based Segregation. It is harder to separate two spoken stories if they both have the same pitch [BN82]. By digitally re-synthesizing speech using LPC analysis, it is possible to hold the pitch of an utterance perfectly constant. It was found that as the fundamentals of two passages were separated in frequency, the number of errors decreased[3]. It was reported that at zero semitones separation, one hears a single auditory stream of garbled, but speech-like sounds, at one half semitone one very clearly hears two voices, and it is possible to switch one's attention from one to the other. Note that a fundamental of 100 Hz was used, and that a half of a semitone (1/12 octave) corresponds to a factor of only 1.03 in frequency. In another experiment, with a fundamental pitch difference of only 2 Hz for a synthesized syllable, virtually all subjects reported that two voices were heard. At a difference of 0 Hz, only one voice was reported.

Harmonics. On a log scale, speech harmonics move up and down in parallel as the pitch of an utterance changes. Harmonics that maintain such a relationship are probably perceived to be related to the same sound source. There is also evidence that supports the idea that changing harmonics can be used to help "trace out the spectral envelope" of the formant frequencies for speech. Two adjacent harmonic peaks can be connected by more than one spectral envelope. However, by analyzing the movement of the peaks as the fundamental changes, it is possible to unambiguously define the formant envelope.

Automatically Recognizing Streams

While this paper focuses on what attributes of the cocktail party effect can be used for enhancing user interfaces that *present* speech information to the user, it is worth considering the recognition problem briefly. It is generally difficult to find tractable and accurate computational solutions to recognition problems that humans find simple (e.g., speech or image comprehension).

> We want to understand the segregation of speech sounds from one another and from other sounds for many practical as well as theoretical reasons. For example, current computer programs that recognize human speech are seriously disrupted if other speech or nonspeech sounds are mixed with the speech that must be recognized. Some attempts have been made to use an evidence-partitioning process that is modeled on the one used by the human auditory system. Although this approach is in its infancy and has not implemented all the heuristics that have been described in the earlier chapters of this book, it has met with some limited success. ([Bre90] page 532)

[3]Note that there was an increase in error rate if the signals were exactly an octave apart.

In 1971, researchers at Bell Labs reported on a signal processing system for separating a speech signal originating at a known location from a background of other sounds [MRY71]. The system used an array of four microphones and simple computational elements to achieve a 3–6 dB noise suppression. This scheme was somewhat impractical, as the source had to remain exactly centered in the microphone array. It was proposed that an ultrasonic transmitter could be carried, so that the system could track the speaker. Recent work in beam-forming signal-seeking microphone arrays appears promising, though much of the effort is geared toward teleconferencing and auditorium environments [FBE90]. With three microphones it is possible to reject interfering speech arriving from non-preferred directions [LM87]

Bregman discusses several systems based primarily on tracking fundamentals for computationally separating speakers (see also [Zis90]). This scheme is somewhat impractical because not all speech sounds are voiced, and the fundamental frequency becomes difficult to track as the number of speakers increases. Weintraub found improvements in speech recognition accuracy in separating a stronger voice from a weaker one [Wei86].

Keep in mind that much of the speech segregation task performed by humans is based in part on knowledge of the transition probabilities between words in a particular context. The use of this technique is feasible for limited domain tasks, but it is unlikely to be computationally tractable for any large domains in the near future.

Stream Segregation Synthesis

There has been a recent surge of work in the area of real-time three-dimensional auditory display systems [Coh90]. This activity has been partially motivated by the availability of inexpensive digital signal processing hardware and the great interest in “virtual environments” and teleoperator systems. A contributing factor has also been advances in understanding of human spatial hearing and computational ability to synthesize head-related transfer functions (HRTFs; directionally sensitive models of the head, body, and pinna transfer functions) [Bla83]. These systems usually rely on the use of stereo headphones, and synthesize sounds that are localized *outside* of the head.

The fundamental idea behind these binaural simulators is that in addition to creating realistic cues such as reflections and amplitude differences, a computational model of the person-specific HRTF simulates an audio world [WWF88]. Multiple sound sources, for example, can be placed at virtual locations allowing a user to move within a simulated acoustical environment. The user can translate, rotate, or tilt their head and receive the same auditory cues as if a physical sound source were present. These systems provide a compelling and realistic experience and may be the basis for a new generation of advanced interfaces. Current research focuses on improving system latency, the time required to create user-specific HRTF models, and in the modeling of room acoustics.

A different approach to the synthesis of auditory streams has been developed by the Integrated Media Architecture Laboratory at Bellcore in the context of a multiperson multimedia teleconferencing system [LPC90]. This “audio windowing” system primarily uses off-the-shelf music processing equipment to synthesize, or enhance, many of the primitive segregation features mentioned in previous sections. Filters, pitch shifters, harmonic enhancers, distortions, reverberations, echos, etc. were used to create “peer” and “hierarchical” relationships among several spoken channels. While the use of these “rock-n-roll” effects may seem extreme, a recent

description of the work discusses the use of "just noticeable effects" that are barely over edge of perceptibility [CL91]. Similar effects are used for "highlighting" pieces of audio to draw one's attention to it.

Unfortunately, the combination of auditory effects needed to generate these relations appears to have been chosen in a somewhat ad hoc manner, and no formal perceptual studies were performed. The work is important, however, in that it has begun to stimulate awareness in the telecommunications and research communities regarding the feasibility of simultaneously presenting multiple streams of speech in a structured manner.

Application Areas

There are a variety of applications that can benefit from the use of a synthetic segregation system, such as multi-party audio teleconferencing. With present conferencing systems there are limitations to the number of participants that can speak simultaneously (usually one), and it is often difficult to identify one speaker from the others. If video is added to such a conferencing environment, it is possible to add spatial audio cues to help disambiguate speakers. For example, if a 2×2 video mosaic is used, the audio for the person in the upper right hand quadrant can be localized in a corresponding spatial location. Such a system could also use other stream segregation effects to enhance the voice of the speaker who has the "floor" at any given instant.

Another emerging application area is speech-only hypermedia [Aro91]. In this context, speech provides navigational input in a hypermedia database among a linked network of voice recordings. It is desirable to present multiple streams of audio information simultaneously, as can easily be done in a graphics-based system, to circumvent the linear, single channel, nature of speech signals. Using techniques described in this paper, it may be possible to enhance the primary speech signal so that it "remains in auditory focus," compared with secondary, or background, channels that are played in parallel. The goal is to keep the speech signals identifiable and differentiable, so that the user can shift attention between the various sound streams. This will allow for a new type of speech-based navigation—the ability to move between "overheard" conversations.

A final area of interest is the use of speech in a handheld computer environment. As with a hypermedia system, one limitation of a small computer (with a tiny, or non-existent, keyboard and display) is navigating though information spaces. Spatial and perceptual streaming cues can help in presenting a high bandwidth information to the user by displaying multiple streams of audio information simultaneously.

The intent of using these perceptual ideas in applications is to help de-clutter the acoustic space of a user interface. However, the incorporation of such techniques does present new problems and challenges. If a user shifts attention to a background stream how is this communicated to a computer? If a full spatial audio system is used, head movements or head gestures (e.g., a glancing nod in the direction of the desired stream) can be used. Otherwise, speech recognition can provide input to the system, but this may be obtrusive in some application environments. If spatial cues are used, should they be in a user- or world-centered coordinate system? Again, this probably depends on the application.

Summary and Conclusions

The percepts that make up the cocktail party problem are complex and intertwined, so a simple closed form solution is not yet practical to embed in speech user interfaces. This paper brings together relevant information from a variety of sources and summarizes a large body of work. Here is a brief summary of components of the effect that may prove to be useful in building interactive speech communication applications:

- Provide spatial continuity within channels
- Provide spatial disparity between channels
- Associate visual images with audio streams
- Provide F0 continuity
- Micromodulate F0 to enhance voices
- Filter streams into separate frequency bands
- Use different voices (synthetic or recorded)
- Pitch shift voices away from each other
- Do not present too much information simultaneously
- Provide a mechanism to "pull" one voice into focus
- Provide enough time for the user to fully fuse streams

It has not yet been determined how perceptual evidence relating to these cues are combined within the brain. More research must be performed to determine the relevant weightings of these effects in different environments, and how these cues work synergistically.

It is unclear how much useful information from background channels can be gleaned while attending to a particular foreground channel. While it has been shown that users can shift attention, what is of particular interest for many applications is in providing cues that suggest it is time to "scan between the channels."

A higher level way to summarize these ideas is: provide as much continuity within a stream as possible, while making them as differentiable from other streams as is practical, without adding so many effects that they are distracting.

Acknowledgements

Kenneth N. Stevens and Lisa Stifelman provided comments on earlier drafts of this paper.
This work was sponsored by Apple Computer and Sun Microsystems.

References

[Aro91] B. Arons. Hyperspeech: Navigating in speech-only hypermedia. In *Hypertext '91*, pages 133–146. ACM, 1991.

[Bla83] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1983.

[BN82] J. P. L. Brokx and S. G. Noteboom. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10:23–36, 1982.

[Bre90] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

[Bro58] D. E. Broadbent. *Perception and Communication*. Pergammon Press, 1958.

[Che53] E. C. Cherry. Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustic Society of America*, 25:975–979, 1953.

[CL91] M. Cohen and L. F. Ludwig. Multidimensional window management. *International Journal of Man/Machine Systems*, 34:319–336, 1991.

[Coh90] E. A. Cohen. Technologies for three dimensional sound presentation and issues in subjective evaluation of the spatial image. In *Audio Engineering Society 89th Convention*, 1990. preprint number 2943.

[CT54] E. C. Cherry and W. K. Taylor. Some further experiments on the recognition of speech, with one and two ears. *Journal of the Acoustic Society of America*, 26:554–559, 1954.

[DC78] N. I. Durlach and H. S. Colburn. Binaural phenomena. In E. C. Carterette and M. P. Friedman, editors, *Hearing*, volume IV of *Handbook of Perception*, chapter 10. Academic Press, 1978.

[FBE90] J. L. Flanagan, D. A. Berkley, and G. W. Elko. Autodirective microphone systems. Preprint of invited paper for a special issue of Acustica honoring Professor G. M. Sessler, 1990.

[Han89] S. Handel. *Listening: An Introduction to the Perception of Auditory Events*. MIT Press, 1989.

[KS83] B. H. Kantowitz and R. D. Sorkin. *Human Factors: Understanding People-System Relationships*. John Wiley and Sons, 1983.

[LM87] H. Liang and N. Malik. Reducing cocktail party noise by adaptive array filtering. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 185–188. IEEE, 1987.

[LPC90] L. Ludwig, N. Pincever, and M. Cohen. Extending the notion of a window system to audio. *IEEE Computer*, 23(8):66–72, August 1990.

[Moo89] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, 3d edition, 1989.

[MRY71] O. M. Mitchell, C. A. Ross, and G. H. Yates. Signal processing for a cocktail party effect. *Journal of the Acoustic Society of America*, 50(2):656–660, 1971.

[Nor76] D. A. Norman. *Memory and Attention*. John Wiley and Sons, 1976.

[SA89] C. Schmandt and B. Arons. Desktop audio. *Unix Review*, October 1989.

[SCW54] W. Spieth, J. F. Curtis, and J. C. Webster. Responding to one of two simultaneous messages. *Journal of the Acoustic Society of America*, 26(1):391–396, May 1954.

[SW63] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, 1963.

[Wei86] M. Weintraub. A computational model for separating two simultaneous talkers. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 81–84. IEEE, 1986.

[WT54] J. C. Webster and P. O. Thompson. Responding both of two overlapping messages. *Journal of the Acoustic Society of America*, 26(1):396–402, May 1954.

[WWF88] E. M. Wenzel, F. L. Wightman, and S. H. Foster. A virtual display system for conveying three-dimensional acoustic information. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, pages 86–90, 1988.

[Zis90] M. A. Zissman. *Co-Channel Talker Interference Suppression.* PhD thesis, Massachusetts Institute of Technology, 1990.

Barry Arons is a doctoral candidate in the Speech Research Group at the Massachusetts Institute of Technology's Media Laboratory. His research interests include highly interactive systems with emphasis on conversational voice communication. He was leader of the Desktop Audio Project and primary architect of the VOX Audio Server at Olivetti Research California, and a member of the technical staff at Hewlett-Packard Laboratories. He holds a BS in Civil Engineering from MIT, and a MS from the Architecture Machine Group at MIT where he co-designed *Phone Slave* and the *Conversational Desktop*.