Chapter 16

The History and Future of CASA

Malcolm Slaney
IBM Almaden Research Center
malcolm@ieee.org

1 INTRODUCTION

In this chapter I briefly review the history and the future of computational auditory scene analysis (CASA). Auditory scene analysis describes the process we use to understand the world around us. Our two ears hear a cacophony of sounds and understand that the periodic tic-toc comes from a clock, the singing voice comes from a radio and the steady hum is coming from the refrigerator.

The field of computational auditory scene analysis crystallized with the publication of Al Bregman's book "Auditory Scene Analysis" (Bregman 1990). The commonly understood goal is to listen to a cacophony of sounds and separate the sounds from the mixture, just as humans do. I would like to argue that this is *not* what people do. In this review, I will describe some progress to date towards modeling human sound separation, and review why this is the wrong direction for those of us interested in modeling human perception. Instead, we should be thinking about sound understanding. This is clearly a much harder problem, but should provide a better model of human sound separation abilities.

In particular, this paper makes two related points. 1) We need to consider a richer model of sound processing in the brain, and 2) human sound separation work should not strive to generate acoustic waveforms of the separated signals. Towards this goal, this paper reviews the use of a correlogram for modeling perception and understanding sounds, the success at inverting the correlogram representation and turning it back into sound, and then summarizes recent work that questions the ability of humans to isolate separate representations of each sound object.

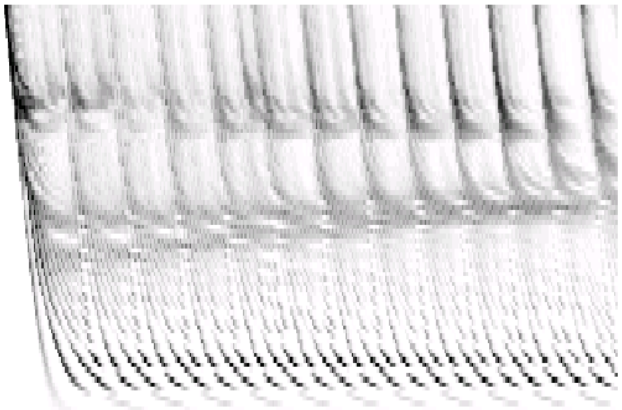

*Figure 16.1.*Simulated auditory nerve firings for the sound "Re". The vertical axis is arranged tonotopically, while time flows horizontally from left to right.

2 AUDITORY MODELING

This section reviews three aspects of auditory modeling: the cochleagram, the correlogram, and their inversion to recover the original sound. This inversion process is interesting, not only because it demonstrates the fidelity of the representation, but also because many sound-separation systems perform their separation in one of these two domains and then want to demonstrate their performance by resynthesizing the cleaned-up sound.

2.1 The Cochleagram

The cochlea transcribes the sounds into a stream of neural firings carried by the auditory nerve. These neural firings are often arranged in order of each nerve’s best frequency to form a cochleagram. The richness of the cochlear data is shown in Figure 16.1. All sound that is perceived travels along the auditory nerve, so it is a complete representation of the perceived sound.

Figure 16.1 shows the output of a cochlear model for the sound “Re.” Horizontal bands are at positions along the basilar membrane where there is significant spectral energy and correspond to the formants that are used to describe speech signals. More interestingly, there is a volley of firings at a regular interval represented by the (slanted) vertical lines. These periodic firings correspond to the energy imparted into the signal by the glottis and their interval directly corresponds to the glottal pitch.

2.2 The Correlogram

The richness and redundancy of the cochleagram suggests the need for an intermediate representation known as the correlogram. The correlogram was first proposed as a model of auditory perception by James Licklider (1951). His goal was to provide a unified model of pitch perception, but the correlogram was also been widely used as a model for the extraction and representation of temporal structure in sound.

The correlogram summarizes the information in the auditory firings using the auto-correlation of each cochlear channel (a channel, in this work, is defined as the firing probabilities of auditory nerves that innervate any one portion of the basilar membrane.) Its goal is collapse and summarize the repetitive temporal patterns shown in Figure 16.1.

The correlogram has met with great success in a number of areas. Meddis and his colleagues (Meddis 1991, Slaney 1990) have demonstrated the ability of a correlogram to model human pitch perception. Assman (1990) and others have shown that the correlogram is a useful representation when modeling the ability of humans to perceive two simultaneous spoken vowels. If anything, models using an ideal correlogram as their internal representations perform even better than humans perform.

The correlogram has served as a compelling visualization tool. In one auditory example created by Steve McAdams and Roger Reynolds, an oboe is split into even and odd harmonics. When the even and odd harmonics are played together at their original frequencies it sounds like the original oboe. But then independent frequency modulation (vibrato) is added to the even and the odd harmonics. The oboe separates into two sounds. The odd harmonics sound like a clarinet because a clarinet has mostly odd harmonic content, while the even harmonics go up an octave in pitch and sound like a soprano. The visual percept is quite striking: two sets of gray dots seem to float independently on the screen. Our task is simply a matter of identifying the dots with the common motion—using whatever tools make sense from a perceptual point of view—synthesizing two partial correlograms, and then resynthesizing the original sound.

2.3 Auditory Inversion

Given the grouping of the sound energy, the final stage is commonly resynthesis, to allow human ears (and funding agencies) hear the separated sounds. Towards this goal, a number of us spent years developing the algorithms that allow us to turn a correlogram movie back into sound (Slaney

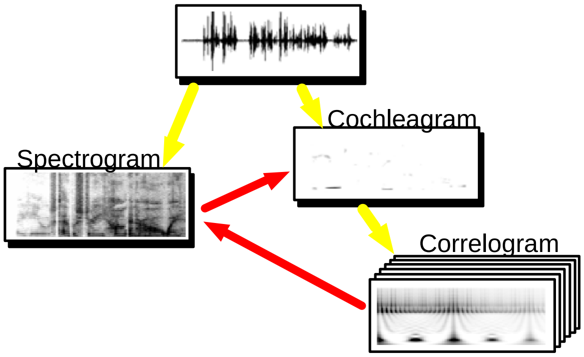

*Figure 16.2.*The correlogram inversion process. From waveform to correlogram and back.

1996). The process requires two steps, as shown in Figure 16.2: 1) cochleagram inversion and 2) correlogram inversion.

From the cochlear representation (the information available on the auditory nerve, or a cochleagram) we invert the loudness compression (undo the automatic gain control in Lyon's model) run the auditory nerve probabilities backwards through the auditory filters, sum the backwards outputs into a single time-domain signal, and then repeat. This procedure, using an idealized cochlear model such as the one produced by Lyon can be done without perceptual differences between the original sound and the sound inversion from the cochleagram.

This procedure was used by Weintraub (1986) in the first real CASA system. He used the correlogram to track the pitch of two different sounds. Each channel was assigned to one sound object or the other, depending on which speaker's pitch was dominant. Then those cochleagram channels from each speaker were grouped, turned back into sound and applied to a speech recognizer. He realized a small improvement in speech recognition performance in a digit identification task.

The second, and more difficult problem, is inverting the correlogram to produce the original cochleagram. The important insight is to realize that each row of the correlogram is a time-varying autocorrelation. An autocorrelation contains the same information as the power spectrum (via an FFT) so each row of the correlogram can be converted into a spectrogram. Spectrogram inversion can be accomplished with an iterative procedure—find the time-domain waveform that has the same spectrogram as the original spectrogram. There is a bit more involved in getting the phase from each cochlear channel to line up, but the result is a sound that sounds pretty close to the original. The inversion problem was solved. This also demonstrates that the correlogram representation is a complete model—little information is lost since the sound from the correlogram is so similar to the original.

3 GROUPING CUES

How does the perceptual system understand an auditory scene? Researchers often think about a number of cues that allow the brain to group pieces of sound that are related (Bregman 1990). But this outlook is inherently a bottom-up approach—only cues in the signal are used to perform grouping. There are other cues that come from a person's experiences, expectations, and general knowledge about sound. This chapter argues that both sets of cues are important for scene analysis. First it is useful to review a range of low-level and high-level grouping cues.

3.1 Low-Level Grouping Cues

Many cues are used by the auditory system to understand an auditory scene. The most important grouping principle is *common fate*. Portions of the sound landscape that share a common fate—whether they start in synchrony or move in a parallel fashion—probably originate from the same (physical) object. Thus the auditory system is well served by grouping sounds that have a common fate.

There are many cues that suggest a common fate. The most important ones are common onsets and common harmonicity. Common onsets are important because many sounds start from a common event and all the spectral components are stimulated at the same time. Common harmonicity is important because periodic sounds—sounds that repeat at a rate between 60 and 5000Hz—have many spectral harmonics that vary in frequency and amplitude in a synchronous fashion. These and other similar cues are well described in Bregman's book.

A different style of low-level cues have been exploited for sound separation by the machine-learning community (Lee 1997). Known as blind-source separation (BSS), these systems generally assume there are N distinct sources that are linearly mixed and received by N microphones. Portions of a signal from the same source are highly correlated and thus should be grouped. BSS relies on the statistical independence of different sources. It forces signals apart that do not share a common fate.

A further refinement is possible in the case of one-microphone source separation (Roweis 2003). The original BSS problem remains—find statistically independent sources that sum to the received signals. One-microphone source separation assumes that for any one position in the spectral-temporal plane, only one speaker at a time is present. The problems becomes a matter of allocating the portions of the received signal so that each source is not only independent, but fits a model of speech. One-microphone source separation uses a model of the speaker, encoded as a vector-quantization model in Roweis' work, to guide the separation.

3.2 High-Level Grouping Cues

Many types of auditory scene analysis can *not* be done using simple low-level perceptual cues. Listeners bring a large body of experiences, expectations, and auditory biases to their auditory scene analysis. If somebody says "firetr..." then the only real question is whether I'm hearing the singular or the plural of firetruck, and even that information can be inferred from the verb that follows. Yet, we perceive we've heard the entire word if there is sufficient evidence (more about this in Section 4.3).

Consider an auditory example I presented at a workshop in Montreal. A long sentence was played to the audience. During the middle of the word "legislature," a section of the speech was removed and replaced by a cough. Most people could not recall when the cough occurred. The cough and the entire speech signal were perceived by listeners that understood the English language as independent auditory objects. But one non-native speaker of English did not know the word "legislature" and thus heard the word "legi-cough-ture." His limited ability to understand English gave his auditory system little reason to predict the word "legislature" was going to come next. This is an example of phonetic restoration (Warren 1970).

A paper titled "A critique of pure audition" (Slaney, 1990) talks about a number of other examples where high-level cues can drive auditory scene analysis. These clues include:

Grouping—Think about a collection of whistles. Would these isolated tones ever be heard as speech? In sine-wave speech three time-varying tone are heard as speech (Remez 1984)

Grouping—A click in an African click language is heard as speech during a spoken utterance, but listeners unfamiliar with that language hear the clicks as instrumental percussion during a song.

Vision affects Audition—In the McGurk effect, a subject's visual perception of the speaker's lips affects their auditory perception. (See Figure 16.3.)

Audition affects Vision—In a simple apparent-motion demonstration, audio clues can cause motion perception in a simple visual display.

Categorical perception—A particular speech waveform is heard as two different vowels depending on the acoustic environment of the preceding sentence (Lagefoged 1989)

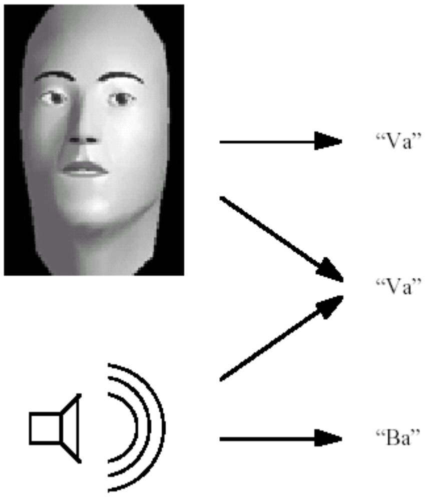

*Figure 16.3.*The McGurk effect. The same auditory stimulus is heard differently when accompanied by video.

In each of these cases, a listener's experience or a completely different input modality affect the sounds we hear. This high-level knowledge is clearly guiding the scene analysis decisions.

3.3 Which is it? Top-down versus bottom-up

Other evidence was described in a paper titled "A Critique of Pure Audition" (Slaney 1996). This paper suggests a number of other effects which call into question a purely bottom-up approach. In the most common cases, common harmonicity causes grouping and we understand the speech. Yet in other cases, speech experience rules the day (i.e. phonetic restoration).

These convoluted connections call into question the hypothesis that sound analysis proceeds in a purely bottom up fashion. More importantly, how is each sound object represented? With high-level expectations and cross-modality input, it seems difficult to believe that each sound object is represented by a neural spike train corresponding to a real (or hallucinated) auditory waveform.

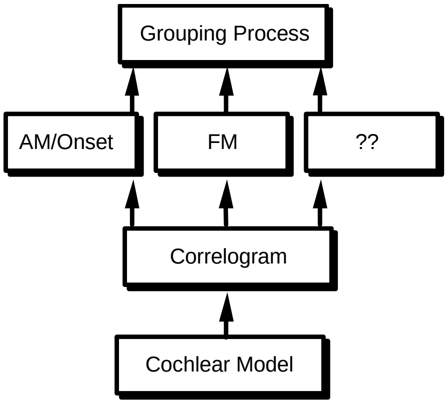

Figure 16.4. The auditory bottom-up grouping process.

4 SOUND SEPARATION MODELS

Researchers use models to concisely describe the behavior of a system. In this section I would like to summarize models that have taken a bottom-up and a top-down view of the world.

Most of the sound separation models to date have evaluated their results based on either the quality of the reconstructions, or the performance of a speech-recognition system on the separate stream outputs. Both approaches largely answer an engineering question: Can we produce a useful auditory scene analysis and improve speech recognition.

The double-vowel perception experiments (e.g. Assman 1990) are a good example of models that attempt to match the human performance data.

4.1 Bottom-Up Models

Much of the original work on scene analysis used a bottom-up approach that was well articulated by David Marr (1982). In Marr's approach simple elements of a scene (either auditory or visual) are group first into simple (2D) cartoons then more sophisticated processing is applied in steps to create a complete understanding of the object (See Figure 16.4). In this model, the brain performs sound separation and object formation, based on all available clues, before performing sound identification.

The first of the bottom-up models for auditory scene analysis was created by Mitch Weintraub, then a Ph.D. student at Stanford University. In Weintraub's work, a correlogram is used to analyze the pitches of two speakers (one male and the other female). The pitches were tracked, thus finding the dominant periodicity for each speaker. Then those (spectral) channels from the original cochleagram with the correct periodicity were inverted to recover an estimate of the original speech signal. His goal was to improve speech digit recognition.

In the 1990's a number of researchers built more sophisticated system to look for more cues and decipher more complicated auditory streams. Cooke and Ellis (2001) wrote an excellent summary of the progress to date using correlograms and related approaches to separate sounds. The essential goal is to identify energy in the signal which shares a common fate, group this energy together, and then resynthesize. Common fate identifies sounds which probably came from the same source, often because the energy in different portions of the signal shares a common pitch, or a common onset.

4.2 Top-Down Models

The bottom-up models use information from the sound to group components and understand an auditory scene. Except for information such as the importance of pitch or onsets, there is little high-level knowledge to guide the scene-analysis process. On the other hand, there is much that language and our expectations tell us about a sound.

Perhaps the best example of a top-down auditory-understanding system is a hidden-Markov model (HMM) based speech-recognition system. In an HMM speech-recognition system, a probabilistic framework is built that calculates the likelihood of any given sequence of words given the acoustic observations. This likelihood calculation is based on low-level acoustic features, often based on an acoustic model known as MFCC (Quatieri 2002), but most of the power in the approach is provided by the language constraints.

The complexity of the language model directly affects the performance of a speech-recognition system. In the simplest example, once the sounds for "firetr" are confidently heard, then the speech recognition system is likely to recognize the utterance as "firetruck" regardless of what sounds are heard next.

The complexity of a speech-recognition system's language model is often described by its perplexity, or the average number of words that can follow any other word. Smaller perplexity means that fewer words can follow, the language is more constrained, and the recognizer's job is easier. In a radiological task, the words are specialized and the perplexity is 20, while in general english the perplexity is 247 (Cole 1996). One of the first commercially successful applications of automatic-speech recognition was for medical transcription, where a relatively small amount of high-level knowledge could be encoded as a low-perplexity grammar.

An even more constrained example is provided by the score-following or music-recognition systems (Pardo 2002). In this case, the system knows what notes are coming and only needs to figure out when they are played. The complexity of a musical signals means that this task can only be accomplished with a very narrow, high-level constraint.

4.3 Mixtures

In practice neither model, top-down or bottom-up, can explain the ability of human listeners to analyze an auditory scene. Clearly low-level cues prime the sound-analysis process—we generally do not hallucinate our entire world. These low-level cues are important, but do not explain simple auditory effects such as phonetic restoration.

Our brains have an amazing ability to hear what might or might not be present in the sound. Consider a slowly rising tone that is interrupted by a short burst of loud noise. As long as the noise burst is loud enough, we perceive a continuous tone that extends through the noise burst, whether the tone is actually present or not. Our brains hear a continuous tone as long as there is evidence (i.e. enough auditory nerve spikes at the correct cochlear channels) that is consistent with the original hypothesis (the tone is present).

This simple demonstration calls into question the location of the hallucinated tone percept. A purely low-level bottom-up model suggests that some portion of the brain has separated out a set of neural spikes that correspond to the phantom tone. The auditory system sees the same set of spikes, with or without the noise burst, and perceives a continuous tone. Instead it seems more likely that somewhere a set of neurons is saying “I hear a tone” and these neurons continue to fire, even when the evidence is weak or non-existent. In other words, object formation is guiding the object segmentation process.

At this point it should be clear that a wealth of information that flows up into the brain from the periphery, and a large body of the listener’s experience and expectations affect how we understand sound. In the middle information and expectations collide in a system that few have tackled. Grossberg’s work (2003) is a notable exception.

The more interesting question, at least for somebody that spent a lot of time developing auditory-inversion ideas, is whether the brain ever assembles a high-fidelity neural coding that represents the pure auditory object. The phonetic restoration illusion only works when the noise signal is loud enough so that the missing speech sound *could* be masked by noise. In other words, the brain is willing to believe the entire word is present as long as it does not violate the perceptual evidence. This is clearly expectation driven, top-down processing. But does the brain represent this missing information as a complete representation of the auditory signal, or use a high-level token to represent the final conclusion (I heard the word “legislature”)?

5 CONCLUSIONS

The purely bottom-up approach to auditory perception is clearly inconsistent with the wealth of evidence suggesting that the neural topology involved in sound understanding is more convoluted. One can build a system that separates sounds based on their cochleagram or correlogram representations, but this appears inconsistent with the functional connections. Instead, our brains seem to abstract sounds, and solve the auditory scene analysis problem using high-level representations of each sound object.

There has been work that addresses some of these problems, but it is solving an engineering problem (how do we separate sounds) instead of building a model of human perception. One such solution is proposed by Barker and his colleagues (2001) and combines a low-level perceptual model with a top-down statistical language model. This is a promising direction for solving the engineering problem (how do we improve speech recognition in the face of noise) but nobody has evaluated the suitability of modeling of human-language perception with a hidden-Markov model.

A bigger problem is understanding at which stage acoustic restoration is performed. It seems unlikely that the brain reconstructs the full acoustic waveform before performing sound recognition. Instead it seems more likely that the sound understanding and sound separation occur in concert and the brain only understands the concepts. Later, upon introspection the full word can be imagined.

Much remains to be done to understand how humans perform sound separation, and to understand where CASA researchers should go. But clearly systems that combine low-level and high-level cues are important.

6 ACKNOWLEDGEMENTS

Many people helped shape the ideas in this chapter. I am especially grateful for the many wonderful discussions and debates we have had at the Stanford CCRMA Hearing Seminar over the past 15 years. I appreciate the time and energy that Richard F. Lyon and John Lazzaro have spent guiding my thinking in this area. Dan Ellis, in particular, had many useful comments about this chapter.

References

Assman, P. F. and Summerfield, Q., 1990, Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies, *J. Acoust. Soc. Am.* 88, pp. 680–697.

Barker, J., Cooke M., and Ellis, D. P. W., 2001, Integrating bottom-up and top-down constraints to achieve robust ASR: The multisource decoder. Presented at the CRAC workshop, Aalborg, Denmark.

Bregman, A. S., 1990, *Auditory Scene Analysis*, MIT Press, Cambridge, MA.

Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V. (eds.), 1996, Survey of the State of the Art in Human Language Technology, http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html.

Cooke, M. and Ellis, D. P. W., 2001, The auditory organization of speech and other sources in listeners and computational models, *Speech Communication*, vol. 35, no. 3–4, pp. 141–177.

Grossberg, S., Govindarajan, K.K., Wyse, L.L., and Cohen, M.A., 2003, ARTSTREAM: A neural network model of auditory scene analysis and source segregation. *Neural Networks*.

Ladefoged, P., 1989, A note on 'Information conveyed by vowels,' *Journal of the Acoustical Society of America,* 85, pp. 2223–2224.

Lee, T.-W., Bell, A., Lambert, R.H., 1997, Blind separation of delayed and convolved sources. In: *Advances in Neural Information Processing Systems,* vol. 9. Cambridge, MA, pp. 758–764.

Licklider, J. C. R., 1951, A duplex theory of pitch perception, *Experientia* 7, pp. 128–134.

Marr, D, 1982, *Vision*, W. H. Freeman and Co.

Meddis, R. and Hewitt, M. J., 1991, Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification, *J. Acoust. Soc. Am*., vol. 89, no. 6, pp. 2866–2882.

Pardo, B. and Birmingham, W., 2002, Improved Score Following for Acoustic Performances International Computer Music Conference 2002, Gothenburg, Sweden.

Remez, R.E., Rubin, P.E., Pisoni, D.B. & Carrell, T.D.,1981, Speech perception without traditional speech cues, *Science*, 212, pp. 947–950.

Quatieri, T. F., 2002, *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice-Hall.

Roweis, Sam T., 2003, Factorial Models and Refiltering for Speech Separation and Denoising, *Proceedings of Eurospeech03* (Geneva, Switzerland), pp. 1009–1012.

Slaney, M. and Lyon, R.F., 1990, A perceptual pitch detector. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*.

Slaney, M., 1996, Pattern Playback in the ’90s, *Advances in Neural Information Processing Systems 7*, Gerald Tesauro, David Touretzky, and Todd Leen (eds.), MIT Press, Cambridge, MA.

Slaney, M., 1998, A critique of pure audition, *Computational Auditory Scene Analysis*, edited by David Rosenthal and Hiroshi G. Okuno, Erlbaum.

Warren, R. M., 1970, Perception restoration of missing speech sounds. *Science*, 167, pp. 393–395.

Weintraub, M., 1986, A computational model for separating two simultaneous talkers. *Proc. of the International Conference on Acoustics, Speech, and Signal Processing '86*., Vol.11, pp. 81–84.