

Auditory Scene Analysis: Computational Models

1. Introduction

Human listeners have a remarkable ability to separate a complex mixture of sounds into discrete sources. The processes underlying this ability have been termed 'auditory scene analysis' (Bregman 1990) (see also *Auditory Scene Analysis*). Recently, an interdisciplinary field known as 'computational auditory scene analysis' (CASA) has emerged which aims to develop computer systems that mimic this aspect of hearing (Rosenthal and Okuno 1998).

Work in CASA is motivated both by a desire to understand the mechanisms of auditory perceptual organization and by a demand for practical sound separation devices. Currently, automatic speech recognizers perform badly in noisy acoustic environments. It is likely that their performance could be improved by integrating CASA with speech recognition technology. Other applications of CASA include hearing prostheses and music analysis.

This article considers three general classes of CASA system and discusses their relative merits. Evaluation techniques for CASA are described and outstanding challenges in the field are identified.

2. Theoretical Influences

CASA has been motivated principally by Bregman's (1990) account of auditory organization. However, it has also drawn inspiration from Marr's (1982) work in the field of machine vision. Marr regarded visual processing as a series of representational transforms, each of which makes explicit some aspect of the preceding representation. Starting with Cooke (1993), several workers in the field of CASA have acknowledged Marr's influence and have developed systems in which representational transforms of the acoustic signal play a key role.

3. Three Classes of CASA System

The stages of processing in a CASA system mirror the conceptual stages of auditory scene analysis (Fig. 1). The input to a CASA system is a digitally recorded mixture of sound sources; following feature extraction, this acoustic mixture is decomposed into parts (segmentation). Subsequently, a grouping mechanism identifies segments that are likely to have arisen from the same sound source, and combines them to form a structure that corresponds to a perceptual stream. A final evaluation stage appraises the quality of sound separation.

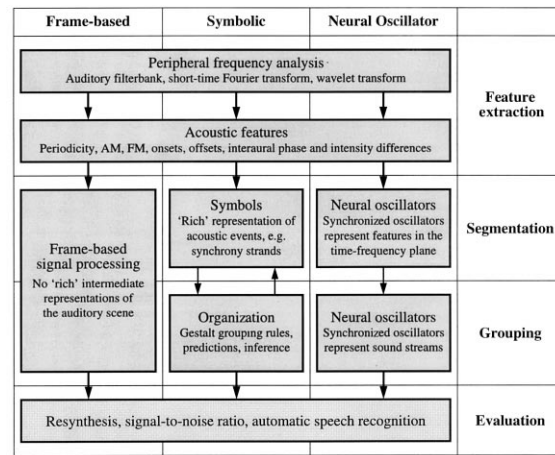


Figure 1

Characteristics of frame-based, symbolic, and neural oscillator approaches to computational auditory scene analysis

Approaches to CASA can be separated broadly into three categories, which differ both in their motivation and in their implementation of the four processing stages identified above. Symbolic systems (e.g., Cooke 1993) adopt a Marrian approach: the segmentation stage derives intermediate representations that characterize properties of the acoustic input. In particular, symbolic systems tend to employ representations of the auditory scene in which temporal continuity has been made explicit. In contrast, frame-based systems (e.g., Denbigh and Zhao 1992) do not use such 'rich' representations; they make grouping decisions within short overlapping time windows ('frames'). Neural oscillator systems (e.g., Wang and Brown 1999) are based more firmly on neurobiological principles than the other types of CASA system. Motivated by reports of neural oscillations in sensory cortex, they achieve segmentation and grouping of acoustic components by forming blocks of synchronized activity in networks of neurons with oscillatory firing responses.

The following sections review the four processing stages of CASA from the perspective of symbolic, frame-based, and neural oscillator systems.

4. Feature Extraction

4.1 Peripheral Auditory Function

The frequency analysis performed by the peripheral auditory system can be modeled by a bank of bandpass filters with overlapping passbands. Each 'auditory filter' channel simulates the frequency response associated with a particular point on the basilar membrane.

Neuromechanical transduction in the cochlea can be modeled by a detailed simulation of inner hair cell function (Brown 1992) or may be approximated by half-wave rectifying the auditory filter outputs and applying a static nonlinearity to compress the dynamic range. CASA approaches which are motivated less strongly by mechanisms of auditory function employ conventional methods of spectral analysis, such as the short-time Fourier transform or discrete wavelet transform (Nakatani and Okuno 1999).

4.2 Auditory Representations of Acoustic Features

Most CASA systems process the peripheral auditory representation in order to derive further representations that make aspects of the acoustic signal explicit. Symbolic and neural oscillator approaches place greater emphasis on these intermediate representations than do frame-based systems.

Human listeners tend to group acoustic components that have the same fundamental frequency (F0). The periodicity of auditory nerve firings provides a cue to F0 and this can be extracted by a running autocorrelation of the activity arising from each auditory filter channel (Brown 1992). Amplitude modulation (AM) in high frequency channels of the auditory filterbank provides another cue to F0 and is used by some CASA systems (Cooke 1993).

Many CASA systems identify the times at which appreciable onsets and offsets of energy occur in the auditory spectrum, since acoustic components which start and stop at the same time are likely to have arisen from the same sound source (Brown 1992). The movement of resonances in the time-frequency plane may also provide information that can be used to track harmonics and formants over time or to identify acoustic components that have a common pattern of frequency modulation (FM) (Mellinger 1992).

Human listeners are also able to separate sounds on the basis of their spatial locations. Phase and intensity differences between the two ears are important cues for spatial location; accordingly, some CASA systems estimate these parameters from a binaural peripheral model (Denbigh and Zhao 1992).

5. Segmentation

The segmentation stage of CASA aims to represent the auditory scene in a manner that is amenable to grouping. Frame-based systems omit this stage of processing: they operate directly on the acoustic features described above.

In many CASA systems, the segmentation stage makes temporal continuity explicit. Typical is the approach of Cooke (1993), which tracks changes in instantaneous frequency and instantaneous amplitude

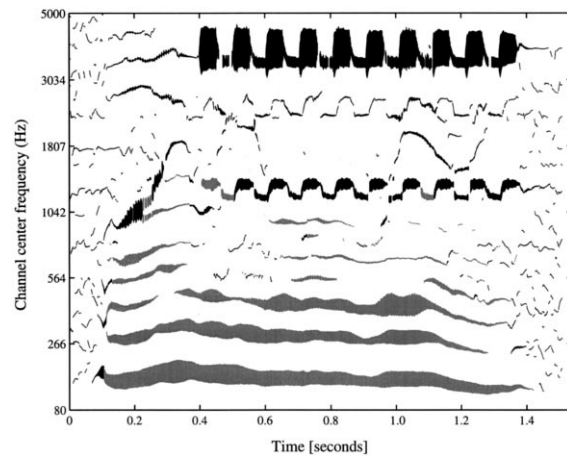


Figure 2

Segregation of speech from an interfering telephone ring in the 'symbolic' CASA system described by Cooke (1993). The utterance is 'why were you all weary?' spoken by a male speaker. A group of harmonically related synchrony strands belonging to the speech source are highlighted in gray. The remaining strands (shown in black) predominantly belong to the telephone sound

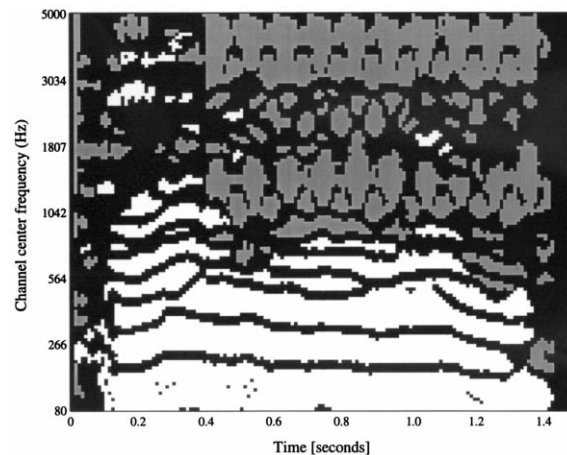


Figure 3

Behavior of the Wang and Brown (1999) system for the sound mixture shown in Fig. 2. Two groups of synchronized oscillators are shown, corresponding to the speech (white) and telephone (gray)

in each channel of an auditory filterbank to create 'synchrony strands' (Fig. 2). Each strand traces the evolution of an acoustic component (such as an harmonic or formant) in the time-frequency plane. This approach offers advantages over frame-based processing: because frame-based schemes make

grouping decisions locally in time, they must resolve ambiguities which would have an obvious solution if temporal continuity were taken into account.

Neural oscillator models of CASA also exploit temporal continuity in the time-frequency plane. In this approach, groups of features that belong to the same acoustic source are represented by a population of neural oscillators whose firing is synchronized. Other groups of features are also represented by synchronized populations but oscillators coding different sound sources are desynchronized. The model of Wang and Brown (1999) employs an architecture consisting of a 'segmentation layer' and a 'grouping layer.' Each layer is a two-dimensional network of oscillators with respect to time and frequency. In the segmentation layer, lateral connections are formed between oscillators on the basis of local similarities in energy and periodicity. Synchronized populations of oscillators emerge that represent contiguous regions in the time-frequency plane (Fig. 3).

6. Grouping

The grouping stage of CASA identifies acoustic components that are likely to have arisen from the same sound source. This is achieved by implementing Gestalt-like grouping principles that underlie auditory organisation in human listeners (Bregman 1990).

A number of CASA systems apply grouping principles using a simple search engine. For example, Cooke's (1993) system initiates search from a 'seed' synchrony strand, and identifies other overlapping strands which match the seed according to some principle of organization. Figure 2 shows a group of synchrony strands that have been identified as harmonically related in this way. A higher-level grouping stage then integrates groups that have been formed by different organizational principles (such as common AM, common FM, onset and offset synchrony).

This approach to grouping is entirely data-driven (i.e., information flows strictly from top to bottom in Fig. 1). However, it is clear that human listeners are able to predict the behavior of acoustic sources (such as speech) in order to disambiguate an acoustic mixture. A more flexible computational framework is needed to combine this top-down information flow with data-driven grouping mechanisms.

Ellis (1999) describes such a 'prediction-driven' architecture. In his system, a world model is updated in response to errors between observed and predicted signals. Context-sensitive behavior emerges as a result of the interaction between model predictions for (possibly overlapping) segments in the auditory scene. For instance, consider a tone that is interrupted by a burst of noise. The observed acoustic evidence may be compatible with both the predictions made for the noise burst and the predictions made for the tone; since the tone is assumed to be continuous rather than intermittent, it is 'perceived' as continuing through the

noise. Hence, Ellis's architecture is able to account for the occurrence of perceptual restoration phenomena (Bregman 1990). A similar 'residue-driven' system is described by Nakatani and Okuno (1999).

In the neural oscillator architecture of Wang and Brown (1999), grouping arises from the dynamics of their oscillator network. In the second layer of the network, excitatory connections are made between segments if the segments are related to the same fundamental frequency and inhibitory connections are made otherwise. As a result, synchronized groups of oscillators emerge which define sound streams (Fig. 3).

7. Evaluation

Most CASA systems allow a time-domain waveform to be resynthesized for each separated 'stream.' Hence, sound separation performance can be judged by listening tests or can be quantified as a change in signal-to-noise ratio (Brown 1992).

A number of workers have evaluated CASA systems as components of automatic speech recognition systems. Early approaches treated CASA as a separate preprocessor; resynthesized speech was presented to an unmodified speech recognition algorithm (Weintraub 1985). However, the performance of this approach was disappointing: although interfering sounds were rejected, the speech was distorted and hence little improvement in recognition rate was obtained. Recent efforts to integrate CASA and speech recognition more tightly have produced encouraging results (Ellis 1999).

8. Future Challenges

Although the CASA systems described here show promise as useful sound separation devices, their performance is still far below that of human listeners. Some workers claim that improved performance can be obtained by an approach which is based more closely on the neurobiology of hearing (Wang and Brown 1999). Equally, however, 'blind' statistical techniques are able to achieve near-perfect sound separation when certain constraints are placed on the statistical properties of the sound sources and the number of sound sensors (Bell and Sejnowski 1995). It remains to be shown, therefore, whether sound separation systems which are strongly motivated by an auditory account have an advantage over engineering approaches.

Another challenge for CASA systems is real-time performance; currently, they tend to be computationally expensive and nonreal-time. The neural oscillator approach may offer some advantage here, since parallel and distributed architectures are easier to implement in hardware than are the complex algorithms associated with symbolic CASA approaches.

See also: Fechnerian Psychophysics; Psychophysics; Psychophysical Theory and Laws, History of; Signal Detection Theory; Signal Detection Theory, History of

Bibliography

- Bell A J, Sejnowski T J 1995 An information maximization approach to blind separation and blind deconvolution. *Neural Computation* **7**: 1129–59
- Bregman A S 1990 *Auditory Scene Analysis*. MIT Press, Cambridge, MA
- Brown G J 1992 *Computational auditory scene analysis: A representational approach*. Ph.D. thesis, University of Sheffield
- Cooke M 1993 *Modelling Auditory Processing and Organisation*. Cambridge University Press, Cambridge, UK
- Denbigh P N, Zhao J 1992 Pitch extraction and the separation of overlapping speech. *Speech Communication* **11**: 119–25
- Ellis D P 1999 Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Communication* **27**: 281–98
- Marr D 1982 *Vision*. W H Freeman, San Francisco
- Mellinger D K 1992 Event formation and separation in musical sound. Ph.D. thesis, Stanford University, CA
- Nakatani T, Okuno H G 1999 Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication* **27**: 209–22
- Rosenthal D F, Okuno H G (eds.) 1998 *Computational Auditory Scene Analysis*. Erlbaum, Mahwah, NJ
- Wang D L, Brown G J 1999 Separation of speech from interfering sounds using oscillatory correlation. *IEEE Transactions on Neural Networks* **10**(3): 684–97
- Weintraub M 1985 A theory and computational model of auditory monaural sound separation. Ph.D. thesis, Stanford University, CA

G. J. Brown

Auditory System

We perceive an acoustic landscape through a system of exquisite microarchitectures along the auditory pathway to the brain. These structures range from the biomechanical at the periphery to topographical organizations more centrally. Rarely stationary, the acoustic landscape is transformed ultimately into our familiar objects of auditory perception such as speech and music. The ocean of sound waves surrounding us encounters two small points of sensory transduction on both sides of the head—leading remarkably to our perception of a full three-dimensional world of sound. The acoustic medium is often viewed as a complementary source of information; after all, higher order primates for the most part are highly visual beings. However, the physical characteristics of sound offer certain advantages in the detection of an object regardless of its location relative to the animal, such as

the ability to detect the location and movement of a sound source in darkness or when the line of sight is occluded. In this article the serial and parallel processing pathways of the auditory system are described briefly, together with observations on the neural coding of auditory information. The term ‘neural code’ refers to a precise language—inferred by experimental observation—that reliably conveys information along neural pathways. Whether the code is actually used by the animal can only be evaluated with respect to the code’s causal impact on the behavior of the animal.

The auditory system, like other sensory systems, consists of a series of connected nuclei that process acoustic information via a complex neural code that necessarily must exist in time and anatomical space (or place). Neural activity in the auditory system is capable of preserving the temporal fine structure of acoustic waveforms, up to a certain degree. Consequently, the timing of individual action potentials likely serves as a major channel of auditory information, particularly at the early levels of the auditory pathway. Centuries ago Jean Baptiste Joseph Fourier (1768–1830) defined the reciprocal relationship between time and frequency of sinusoidal waveforms that form a composite known as the spectrum. A single sinusoidal waveform can be represented as a periodic function in the time domain, but also represented as a single component among others in the composite spectrum of frequencies. This dual relationship plays a prominent role in the representation of acoustic information in the auditory system, beginning with the anatomical frequency mapping on the basilar membrane of the cochlea. The majority of nuclei in the auditory system reflect this frequency mapping, an observation that is commonly referred to as cochleotopic or tonotopic organization.

Acoustic vibrations are initially transferred to the tympanic membrane through the external ear, which causes an increase in sound pressure due to resonance formed in the ear canal. The familiar fleshy appendages we commonly refer to as ‘ears’ are called pinnae. Pinnae play an important role in our ability to locate sounds in space. Next, the middle ear couples acoustic energy from the external ear via three small bones called the malleus, incus, and stapes to the oval window of the cochlea in the inner ear.

1. Cochlea

The cochlea forms the foundation for signal processing in the auditory periphery (see Geisler 1998). It is embedded deep in the temporal bone and contains a coiled basilar membrane that is shown unwrapped in Fig. 1. Three fluid-filled ducts can be observed (scala vestibuli, scala media, and scala tympani), divided by a stiff cochlear partition composed of the basilar membrane and associated structures. The organ of Corti sits atop the basilar membrane and contains the