Auditory Scene Analysis: Computational Models

Martin Cooke, University of the Basque Country, Vitoria, Spain

© 2015 Elsevier Ltd. All rights reserved.
This article is a replacement of the previous edition article by G.J. Brown, volume 2, pp. 942–946, © 2001, Elsevier Ltd.

Abstract

Listeners have to make sense of a complex acoustic world containing overlapping sound sources that must be organized into individual auditory objects. Computational auditory scene analysis concerns the use of algorithms inspired by human sound perception whose aim is to extract properties of constituent sound sources in a complex mixture. Starting with representations based on models of how sound is processed in the peripheral auditory system, typical computational auditory scene analysis techniques function by decomposing the mixture into components followed by selective recomposition into groups of components that appear to emanate from a single source. Grouping processes can be informed by information in the signal itself or by the use of prior statistical models of sound sources. This article outlines some of the principal signal decompositions used in models of auditory grouping and goes on to describe a decoder that combines both signal- and model-driven grouping processes.

Introduction

We live in an acoustically cluttered world and are often forced to pick out the sound source of interest from the mixture reaching our ears. Computational auditory scene analysis (CASA) is the branch of machine perception that attempts to extract information about the location, identity, and content of individual acoustic sources in a mixture (Lyon, 1983; Brown and Cooke, 1994; Ellis, 1999; Wang and Brown, 2006). CASA algorithms differ from other engineering approaches to source separation by being both inspired and constrained by human perceptual organization of sound. In practice, this means that CASA algorithms operate with signals from one or, at most, two microphones; employ a first stage of signal analysis using a computational model of the peripheral auditory system; and embody auditory grouping processes informed by auditory scene analysis.

CASA grew in part from investigations of the perceptual organization of sound in listeners (Bregman, 1990; Darwin, 1891), although early research in the nascent field of CASA (e.g., Cooke, 1993) was also influenced by computer vision, especially the work of Marr (1982). Computational techniques in hearing and vision distinguish between bottom-up ‘signal-driven’ processes that extract information from sensor data, and top-down ‘model-driven’ processes that use prior knowledge of objects to interpret features derived from the signal. However, there are important differences between the two modalities that are reflected in computational approaches in each field. While vision systems routinely handle large numbers of spatially separated objects, the number of individual acoustic sources of interest is usually far smaller, but the separation problem is more challenging since information about each source is overlapped in both time and frequency. Likewise, while attentional-controlled searches for salient objects in the spatial dimension comprise a key aspect of computer vision systems, temporal processing plays a more important role in accessing information in computational audition (e.g., following a conversation in the presence of other voices or tracking a single instrument in a polyphonic composition).

As described in the sections of this article, CASA algorithms operate by first decomposing the incoming sound mixture in order to transform the signal into domains where the constituent sources are more easily separable, then selectively recombinining signal elements, often based on auditory grouping principles, where there is evidence that the elements originated from the same acoustic source.

Signal Decomposition

The decomposition stage converts one or two one-dimensional microphone signals into one or more two- or three-dimensional representations of sound, whose purpose is to more clearly reveal features of the individual acoustic sources such as periodicity or spatial location, thereby facilitating sound separation. Typically, decomposition proceeds by converting each microphone signal into an ‘auditory spectrogram’, a representation of auditory nerve activity as a function of frequency and time. Example auditory spectrograms are shown in Figure 1 for a female talker, a male talker, and a mixture of the two. Each image results from passing the signal through a bank of 100 bandpass filters whose shapes and center frequencies are first approximations to human auditory filter shapes derived from psychophysical studies (Glasberg and Moore, 1990). The temporal envelope in each frequency band has been extracted, smoothed, logarithmically compressed, and converted to a grayscale, in which darker areas correspond to more energetic sound components.

While the auditory spectrogram is primal in documenting the relative energy balance of the constituent sources in a mixture at each time and frequency point, fine timing information is lost in its construction, and other transformations are typically applied to the signal at the high temporal resolution output of each modeled auditory nerve fiber. The two most common transformations are based on auto- and cross-correlation, inspired by early physiological models of Licklider (1951) and Jeffress (1948), respectively. The ‘correlogram’ representation, shown in Figure 2(a), is the
collection of autocorrelation functions of auditory filter outputs, applied independently in each frequency band. In principle, the correlogram is a three-dimensional representation of sound, with dimensions of cochleotopic frequency, autocorrelation delay (‘lag’), and time, but in practice it is computed on short overlapping segments of the signal. Autocorrelation enhances periodicities in each frequency region, and, as explained in this article, it is the across-frequency grouping of common periodicities that can be used to separate sound sources at those time instants where one or both are periodic, as is the case for speech signals at moments where the vocal cords are active, or for plucked instruments.

In a similar manner, pairs of microphone signals can be cross-correlated to highlight the delays at which the two signals are most similar, based on interaural time difference cues to the source azimuth. More recently, for sensors mounted on head simulators, CASA models (e.g., May et al., 2011) have attempted to reinforce time difference cues in the cross-correlogram with level difference estimates motivated by the observation that energy received at the sensor on the opposite side of the head to the source is substantially attenuated relative to that of the closest microphone. Other low-level signal transformations applied in CASA systems include the extraction of temporal envelope modulations, which can encode the fundamental frequency of periodic sources and other modulation components (e.g., Hu and Wang, 2004), as well as source onset and offset enhancement (e.g., Brown and Cooke, 1994).

**Grouping: Recovering Individual Sound Sources by Selective Recombination**

Having transformed the acoustic mixture into representations in which evidence for each individual source is more salient, the remaining task for CASA algorithms is to determine which elements belong together in order to reconstruct one or more sources. Early algorithms focused exclusively on the application of auditory grouping principles based on information in the signal itself. More recent approaches have incorporated prior knowledge of sound sources, often embodied in probabilistic models. We consider each in turn.

**Signal-Driven Grouping**

Figure 1 exposes a multiplicity of cues that might be used to recombine signal elements into source descriptions without prior knowledge of the individual sources. Considering the lower panel which depicts the mixture of male and female...
voices, it is evident that each voice can be associated with a number of time-varying partials that are harmonics (integer multiples) of a common fundamental frequency. These components, apparent in the low- and mid-frequency regions, share a common onset time, maintain harmonicity as they change smoothly over time, and to a large extent possess simultaneous offsets. In the mid- and high-frequency regions, we see evidence of frequency components enhanced by the resonant cavities of the vocal tract. Again, these share common onsets and offsets with the harmonic partials and change smoothly over time. A closer look reveals periodic vertical structures, which are most evident around 0.9 s for the male voice. These correspond to amplitude modulations caused by the interaction of more than one harmonic component in a single auditory filter, and for midfrequencies the modulation period is equivalent to the harmonic separation (i.e., the fundamental frequency). In principle, then, periodically excited portions of a sound source such as speech can be recombined on the basis of a common fundamental frequency along with continuity constraints and onset–offset synchrony.

A representative signal-driven grouping algorithm based on fundamental frequency (Ma et al., 2007) is depicted in Figure 2. Panel A shows the correlogram for a single frame of a mixed male and female talker signal. The correlogram represents the periodic structure of the signal in each frequency region in a small time interval. The main value of the underlying autocorrelation function comes from its property of preserving periodicities in the input signal while removing phase information. In response to a periodic signal, each frequency region of the correlogram contains peaks at delays corresponding to multiples of the frequency of dominant components in that frequency region. So, for a harmonic series, low-frequency channels will produce a peak at the frequency of the lowest component, while progressively higher channels will possess peaks at the frequency of the nearest component and at integer multiples of that frequency. For example, in response to a voice with a fundamental frequency of f Hz, peaks will be present at delays of 1/f s, 2/f s, and so on in the low-frequency region; at delays of 1/2f s, 2/2f s, and so on in a slightly higher frequency region; and at lags of 1/3f s, 2/3f s, 3/3f s, and so on for the next frequency region. Critically, note that each frequency region contains a correlogram peak at a delay of 1/f s. It is this coincidence that makes the correlogram a powerful representation for discovering fundamental frequencies of constituents of a mixture. The coincident delays can be seen as vertical structures in Figure 2(a). In order to extract dominant fundamentals, it is usual to sum
the correlogram across frequency channels, resulting in a summary correlogram as depicted at the top of Figure 2(b). The largest peak corresponds to the female voice at a lag of around 4 ms (250 Hz), while other peaks can be seen at around 8, 9, and 12 ms. In principle, it is possible to identify which frequency regions contributed to each peak, and thereby group together those that appear to come from the same source. Unfortunately, due to the presence of large numbers of autocorrelation peaks in the midfrequencies, it is not feasible in practice to use the correlogram directly in this way to perform source separation. One solution, adopted by Ma et al. (2007), is to sharpen the correlogram using Gabor cosine operators in what is essentially a matched filter operation taking advantage of the local ‘pyramidal’ structure of the correlogram. This results in a less ambiguous summary correlogram, but peaks in the summary are still not guaranteed to correspond to the fundamental frequency of one of the constituent sources, as shown by the estimates derived in this way in the third panel of Figure 2(c). Consequently, a multipitch tracking process is normally used to exploit temporal continuity constraints in order to resolve ambiguous local pitch estimates.

Grouping using intrinsic signal properties such as common fundamentals can go some way to reconstructing an individual sound source. However, there are limits as to what the signal-driven approach can accomplish. One specific issue concerns those parts of the signal that are not periodically excited, such as the voiceless consonants /s/, /th/, and /f/ in ‘set,’ ‘three,’ and ‘four’ in Figure 1. Grouping of voiced and voiceless elements has received relatively little attention in CASA (but see Hu and Wang, 2011), but it can in principle make use of the onset–offset synchronies that result from an alternation of voiced and voiceless elements, or interaural time and level differences that signal a common source azimuth (Harding et al., 2006).

**Pure Model-Driven Grouping**

Individual acoustic sources are often highly structured, containing regularities that can be captured by discrete or probabilistic models. For instance, in the case of speech, constraints exist at every level, from articulatory limitations (e.g., vocal tract size or vocal fold vibration rate) through to linguistic determinants of lexical, grammatical, and semantic possibilities. Similar considerations apply to musical instruments and animal vocalizations. Even less-structured sounds, including those originating from the combination of multiple sources (e.g., babble, traffic, leaves rustling, and waves crashing), exhibit patterning in their spectral profile or temporal modulations. In all these cases, source models can be learned from data, and more recent CASA systems have shown how this kind of prior knowledge can be used to separate a mixture into its individual components. The key idea is to combine statistical models for each source believed to be present at each instant in the mixture in an attempt to explain the acoustic observations. In this way, sources are separated essentially as a by-product of recognition. The original scheme, proposed by Varga and Moore (1990), has since been extended to cope with the computationally complex process of evaluating all possible combinations of a large number of source models. Source model combination can be extremely effective. Indeed, algorithms have been demonstrated that outperform human listeners in limited-domain tasks such as the recognition of speech from one talker in the presence of a competing talker (Rennie et al., 2010).

**Identifying a Target Source in an Unknown Background**

The fundamental problem faced by listeners is to extract information from a single target sound source in the face of other sounds, whose composition (e.g., type and number) may vary over time. Pure model-driven grouping fares less well in situations where the number of active sources has to be estimated, and also requires statistical models at some level of structural detail for all sources present. A source separation system aimed at tracking a single target in an arbitrary background was proposed by Barker et al. (2005). The system makes use of both signal-driven and source model-driven grouping, and is based on the notion of ‘glimpses,’ motivated by energetic masking considerations in the auditory periphery. As a consequence of loglike amplitude compression in the cochlea, auditory nerve activity in each frequency region is dominated by the most energetic source at that instant to such an extent that weaker sound sources make almost no impact on activity unless they are within a few decibels of the stronger source. Because many sounds exhibit sparse concentrations of energy across time and frequency, the likelihood of any pair of sources contributing similar energies in any spectro-temporal location is quite low. In other words, while sources are acoustically transparent, they are better characterized as opaque after encoding in the auditory periphery. The mixture shown in Figure 1 illustrates the opacity approximation. Locally, each part of the auditory spectrogram is nearly identical to the corresponding region of the auditory spectrogram for one of the constituent talkers. Some interactions between the sources are visible (e.g., in the low-frequency region at around 0.6 s), but most of the spectro-temporal plane provides a relatively undistorted view of one or other sources.

The glimpsing idea is further illustrated in the upper panel of Figure 3, which uses a form of color coding to denote spectro-temporal regions in a mixture of two talkers that is dominated by one or the other talker. Each compact region of a single color is a potential glimpse of one source. Of course, glimpses do not arrive at the auditory system with a convenient color code, and the computational task is akin to taking two or more jigsaws, throwing away some large fraction of the pieces of each, and attempting to identify the original pictures. Recast in this manner, the focus of glimpse-based CASA is on the twin problems of glimpse detection and integration. Detection is well suited to signal-driven grouping. Potential glimpses of sufficient spectro-temporal extent typically contain cues based on features such as periodicity or onset synchrony which allows them to be constructed using the kinds of signal-driven grouping algorithms outlined in this article. Glimpse detection avoids the need for extensive grouping of elements, especially across time, and thereby greatly simplifies the CASA problem.

The glimpse integration stage requires probabilistic models for the target sound source. For speech, these would typically be hidden Markov models for phonemes, words, or some other unit. The goal, shown in the panel in Figure 3, is to find both
the most likely word sequence and the set of glimpses that explain the acoustic observations. Decoders of this type rely on missing data techniques (Cooke et al., 2001; see also Seltzer et al., 2004) to permit models to be evaluated against the partial information contained in glimpses.

A brute force approach to glimpse integration would require the evaluation of all possible subsets of glimpses against all permissible model sequences, something that is clearly infeasible. However, a computationally tractable algorithm for glimpse integration is available, as illustrated in Figure 3. Consider first the way that a conventional decoder for sources such as speech operates. At each moment in time, the decoder maintains a very large number of alternative partial interpretations of the data (e.g., incomplete phrases) up to that point, each of which has an associated likelihood. At the next time instant, a new acoustic observation becomes available and is used to update all existing interpretations. Finally, on reaching the end of the utterance, the most likely interpretation is chosen as the recognized result. The glimpse decoder operates in a similar fashion up to the point at which a new glimpse is encountered. At this point, each ongoing hypothesis is cloned. In subsequent time instants, one hypothesis is evaluated by including the glimpse in its interpretation, while its cloned counterpart excludes it. When the glimpse ends, all pairs of cloned hypotheses are compared and the most likely path of the pair is retained, under the assumption that later-arriving glimpses will not affect the earlier interpretation. Compared to a conventional single-source decoder, the complexity of the glimpse decoder is increased by a factor that in the worst case is 2 to the power of the largest number of temporally overlapping glimpses encountered. This number depends in turn on the power of signal-driven grouping in detecting glimpses. In this way, the decoder trades off signal and model-driven grouping: the more successful the signal-driven grouping stage, the fewer alternatives are presented to the model-driven stage.

Applications, Evaluations, and Future Challenges

The ability to describe the content of complex acoustic environments will find widespread application in many fields, including human–computer interfaces, health, security, entertainment, and robotics. Key beneficiaries of general-purpose CASA will be hearing-impaired individuals, for whom selective enhancement of sources of interest will be possible. Other applications include robust automatic speech recognition for everyday (i.e., potential noisy) conditions, remote audio surveillance, automated music analysis, transcription, and repair, as well as providing audio descriptions and audio awareness for robots.

Much effort in CASA to date has been dedicated to tackling the ‘cocktail party problem,’ which involves separating or recognizing a voice of interest in a background of one or more distracting voices. This formed the challenge problem for the first global evaluation of speech separation and recognition systems, which took place in 2006 (SSC website). While speech separation in the cocktail party task is an important goal, and one that is particularly relevant for hearing-impaired listeners, it is a largely artificial problem. Realistic sound environments have additional complications such as the presence of sound sources whose type and number vary over time. This scenario forms the basis for current evaluations (CHIME website).

Future systems can be expected to handle moving sources. To counteract these and other dynamic elements of auditory
scenes, CASA systems may well utilize active sensors to simplify the tracking of moving targets while attenuating the contribution of competitors. More sophisticated scenarios demand models of auditory attention capable of focusing on sources of interest while responding to relevant novel events. Richer characterizations of the auditory scene will be possible by using algorithms that construct more complete spatial descriptions than is currently the case, going beyond azimuth estimates to include elevation and distance. An additional challenge is to develop more robust scene analysis algorithms that are capable of handling and indeed exploiting reverberant energy and other convolutional distortions. Finally, it is likely that auditory and visual scene analyses will be more closely integrated in future in order to exploit the advantages of each modality.

See also: Speech Perception; Speech Recognition and Production by Machines; Statistical Pattern Recognition.

Bibliography


Relevant Websites

http://spandh.dcs.shef.ac.uk/chime_challenge/ – The CHiME Challenge (CHiME).