Chapter 5

Computational Auditory Scene Analysis

5.1. Introduction

Until recently, the study of auditory processes has been mainly focused on perceptual qualities such as the pitch, loudness or timbre of a sound produced by a *single source*. Experimentations in psychoacoustics have brought to the fore a relationship between the physical properties of a sound and the feelings which it conveys; the physiological mechanisms underlying this relationship have been sketched. Models of auditory processes have been designed, which are based on the acoustic wave or its spectrum.

Unfortunately, the audio sources around us are rarely active in isolation. We live in a cacophony of superimposed voices, sounds and noise, the resulting spectrum of which is completely different from that of a single source. Each of our ears receives sound waves originating from a multitude of sources. However, it is generally possible to focus one's attention on a specific source and perceive its loudness, its pitch, its timbre, and understand what is being said (when dealing with a speech source), in spite of the presence of competing sounds. Conventional models, which have been designed for isolated sounds, are not powerful enough to account for the perception of multiple sources.

In his day, Helmholtz wondered how we could perceive individual qualities of instruments when they were played together [HEL 77]. But it is with the works of Bregman that *auditory scene analysis* (or ASA) became a research topic in itself [BRE 90]. For Bregman, the study of the emergence of subjective sources (or

---

Chapter written by Alain DE CHEVEIGNÉ.

streams) is central, as it logically comes before the determination of their individual qualities. Bregman's ASA is a transposition of visual scene analysis principles into the field of audition.

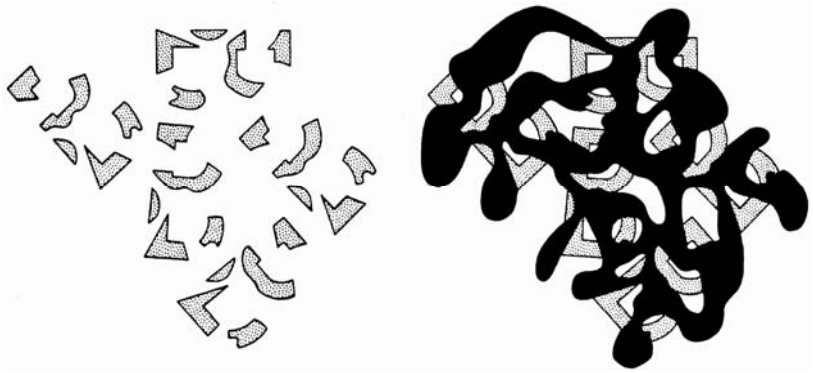


Figure 5.1. *Visual scene analysis. On the left, the fragments are not organized. On the right, the presence of a masking shape enables their grouping on a perceptual basis. ASA searches for similar principles for organizing the auditory world (after [BRE 90])*

Along with the development of computer science and artificial intelligence, attempts have been made to develop *computational auditory scene analysis* (CASA) [BRO 92a, COO 91, ELL 96, LYO 83, MEL 91, WAN 95, WEI 85]. CASA models have a twofold ambition: 1) to describe and explain perceptual processes and 2) to solve practical problems, such as noise removal in a speech recognition system. The influence of computational vision research, in particular the work of Marr [MAR 82] has played a key role.

The concept of the *CASA model* suffers from some ambiguity. For the modeling of perceptual processes, it is not easy to define the border between CASA models and other models, as computational modeling has become quite traditional in several domains. Viewed as a signal processing approach, the specificities and advantages of CASA modeling over other techniques are not completely straightforward. By aiming at being a good auditory model *and* a useful approach, the CASA model runs the risk of reaching neither of these two objectives. Nevertheless, the CASA approach can be fruitful, provided the difference is clearly made between the model and the method, especially when they are evaluated. The insistence in designing a full (and therefore complex) system is a good remedy against the potential reductionism of psychoacoustic models. From a practical viewpoint, applications such as speech recognition need to replicate the noise tolerance ability of the human auditory system. Interesting developments have recently been coming from the CASA approach, in particular the *missing feature theory* [COO 94, COO 97, LIP 97, MOR 98].

5.2. Principles of auditory scene analysis

5.2.1. *Fusion versus segregation: choosing a representation*

In the framework of ASA, the notions of *fusion* and *separation* are often used. Fusion corresponds to situations when some features are attributed to the same audio source (or stream) whereas segregation happens when they are distributed over several sources. To give full meanings to these terms, an internal representation composed of auditory cues must be hypothesized, in which the cues from the various sources can be separated from one another. This may mean a representation of the physical stimulus in the time domain, in the frequency domain or in any one of the various time-frequency representations. Alternatively, a physiological representation can be considered (cochlea-based filter banks, neural coincidence networks, etc.), from which the auditory system is able to extract elements pertaining to each source.

In fact, psychoacousticians use a third type of representation when they describe a stimulus in terms of synthesis parameters (duration, amplitude, frequency or instantaneous phase for each component). This is not exactly a time-frequency representation in the conventional sense, as no representation of this type can provide such an accurate description on the time and frequency axes simultaneously. For example, let us consider a stimulus composed of several sine functions modulated in frequency. In the synthesis operation, the instantaneous frequency is perfectly specified but there is no general method for retrieving these parameters from the stimulus. A time-frequency analysis may provide an approximate estimation, but not a unique exact value that would correspond precisely to the idealistic description of psychoacoustics.

This causes considerable confusion. The *principles of ASA* have been stated by psychoacousticians in terms of synthesis parameters. On the other hand, the CASA model does not have access to this idealistic representation and must deal with what can be extracted from the signal. Many “good ideas” in terms of idealistic representations go flat when they are applied in practice. Revealing these difficulties is one of the merits of the CASA approach.

5.2.2. *Features for simultaneous fusion*

While keeping in mind the abovementioned restrictions, let us consider a stimulus “formed” of a given number of components. We could expect that the auditory system attributes them to a single source, as a sonometer or a speech recognition system would do. Our experience shows that this is not always the case: in general, we “separate the components” of the stimulus and attribute part of them

to each source. The following question thus arises: as, in some cases, the components from distinct sources are separable (by segregation), why are they sometimes perceived as being grouped (fusion)? Fusion and segregation are two sides of a same coin. What are the acoustic features that trigger either one?

Harmonicity. A harmonic relationship between components favors their fusion. This is the case when the stimulus is periodic (voiced speech, some musical instruments, etc.). On the contrary, when the stimulus is inharmonic ("polyperiodicity" [MAR 91]), it seems to contain several sources. Concurrent vowels or voices are easier to understand if they follow distinct harmonic series, i.e. if their fundamental frequencies F_0 are different.

Envelope coherence, attack synchronicity. If individual components start simultaneously and their amplitude varies coherently, they tend to be fused. Conversely, an attack asynchrony favors segregation. This is an example of the more general principle known as *common fate*.

Binaural correlation. If the components of a source all have the same binaural relationship, their fusion is favored. A difference in the binaural relationship between a target sound and a masking sound favors the perception of the target sound.

Coherent frequency modulation. This is another example of the common fate principle. If the time-frequency representation is viewed as a picture, the components with coherent modulation should form a pattern and emerge from the static components or from those with incoherent modulation.

All these features have been proposed and implemented with more or less success in some CASA systems.

5.2.3. *Features for sequential fusion*

Similarly to simultaneous fusion, we could imagine that sounds which follow one another over time are always attributed to the same source (fusion). This is not the case: in some situations, the auditory system divides a sequence of sounds into several distinct streams (segregation). Each stream then seems to evolve independently. Each of them can be chosen and "isolated" by attention. The order of the sounds within a given stream can be distinguished, but not from one stream to another. This phenomenon is exploited in Bach's fugues to create the illusion of several melodic lines with a single instrument. Among the features which determine fusion and segregation, let us mention:

– frequency proximity: a sequence of pure sounds with close frequencies tends to fuse into a single stream. The sounds form distinct streams if frequencies are far apart;

– repetitiveness: segregation trends are reinforced by the duration and the repetitiveness of stimuli;

– repetition rate: presenting sound sequences at a fast rhythm favors segregation. A slower rhythm favors fusion;

– timbre similarity: a sequence of sounds with a common timbre tends to favor fusion. Sounds with very different timbres are less likely to fuse and it is difficult to determine their time order.

On the basis of this enumeration, we could expect that, because of its many discontinuities in amplitude, timbre, etc., speech is not perceived as a coherent stream. Paradoxically, this is not the case: a voice keeps its coherence despite these discontinuities.

5.2.4. *Schemes*

The previously mentioned features depend on the signal and underlie what is called *primitive* fusion. The corresponding mechanisms are automatic and unintentional. They do not depend on any training or cognitive context. Situations also exist where the fusion is based on learned *schemes*, on abstract regularities or on the subject's state of mind. The distinction between primitive versus scheme-based fusion is to be put in parallel with the *bottom-up* versus *top-down* processes in artificial intelligence.

5.2.5. *Illusion of continuity, phonemic restoration*

When a short noise is superimposed upon a continuous tone, the tone seems to continue "behind" the noise. The same impression is observed even if the tone is interrupted during the noise, provided the latter is loud enough. This is called the *illusion of continuity*. Similarly, with speech, if a phoneme is replaced by a relatively loud noise, the missing phoneme is perceived as if it were still present. This is the phonemic restoration phenomenon. The "restored" phoneme can vary according to the context (for instance, a stimulus such as "*eel" becomes "wheel", "peel", "meal", etc., according to the semantic context). Quite surprisingly, it is almost impossible to tell which of the restored phrase's phonemes was missing.

5.3. CASA principles

5.3.1. *Design of a representation*

The analogy with visual scene analysis, on which ASA is based, assumes the existence of a "representation", the richness of which is comparable to the 3D space for objects or the 2D space for images (Marr uses the term 2½D to designate the enriched representation stemming from binocular vision and other perception mechanisms related to depth [MAR 82]). As the acoustic wave is of low dimensionality, the CASA model starts by synthesizing an enriched representation.

5.3.1.1. *Cochlear filter*

The conventional CASA model starts with a filter bank. In principle, these filters conform to what is known on cochlear filtering. In practice, a wide variety of filter banks has been proposed, according to the priority decided by the designer on the proximity to a physical model of the cochlea, the conformity with physiological recordings or psychological data, the ease of implementation, etc. Currently, the most popular filter is the *gammatone* filter, which is relatively realistic and easy to implement [COO 91, HOL 88, PAT 92, SLA 93]. The filters of a cochlear filterbank model are generally of constant width (in Hz) up to 1 kHz. Above this frequency, their width is proportional to their central frequency. An additional delay may be added to the output of the channels so as to compensate for the group-delay differences, and "align" their impulse responses.

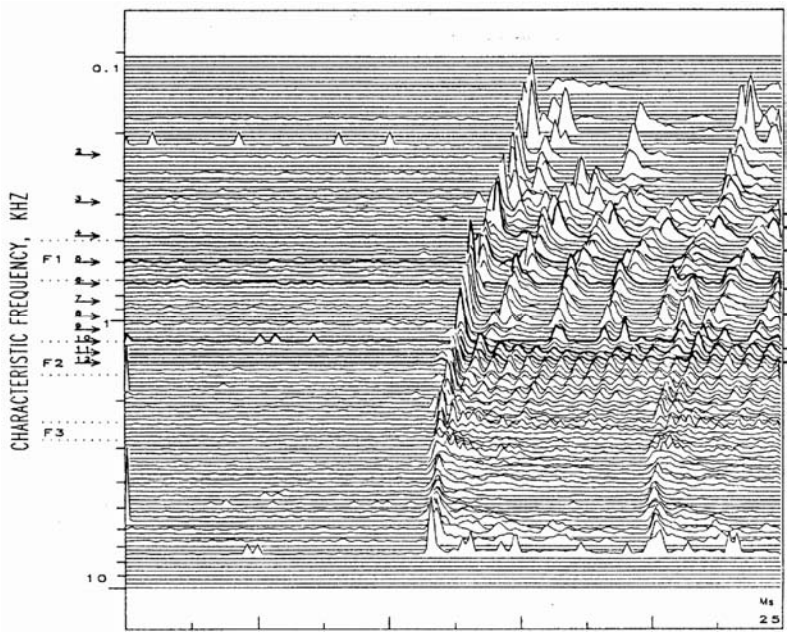


Figure 5.2. *Activity of a group of auditory nerve fibers, for a cat, in response to the synthetic syllable [da]. Cochlear filtering and transduction models try to reproduce this type of response. The progressive delay for low frequency channels (top) due to the propagation time in the cochlea is usually compensated for in the model (after [SHA 85])*

5.3.1.2. *Transduction*

The mechanical vibration of the basilar membrane governs the *probability* of discharge of the auditory nerve fibers which synapse to the inner hair cells.

This process may be modeled with varying degrees of realism:

– since a probability is positive, the transduction shows properties similar to the one of a half-wave rectifier;

– transduction also has compressive properties which can be modeled by a simple instantaneous non-linearity (logarithm, cubic root, etc.), or by an adaptive mechanism: automatic gain control [HOL 90, LYO 82, LYO 84, PAT 92, SEN 85] or hair cell model [MED 86, MED 88];

– in the models by Lyon [LYO 82, LYO 84] and Holdsworth [HOL 90] , the gain for each channel varies according to the activity within a time-frequency neighborhood. The physiological basis of this process is unclear, but it has the

favorable effect of reinforcing the contrast of the representation in the frequency domain (this is an example of confusion between a model and a method). Some other models go further and incorporate an explicit mechanism of spectral and/or temporal differentiation, an example of which is the LIN (*lateral inhibitory network*) of Shamma [SHA 85];

– non-linear transduction is usually followed by low-pass filtering (time-domain smoothing). Depending on the model, this filtering operation is either light (small time-constant) and represents the loss of synchronization observed physiologically in the high frequencies (between 1 and 5 kHz), or more severe, so as to eliminate the periodic structure of voiced speech and to obtain a stable spectrum over time.

The output of the filter/transduction module can be seen either as a sequence of short-term spectra, or as a set of parallel channels each carrying a filtered version of the signal. This is a high-dimensional representation which is a first step towards a favorable substrate for scene analysis.

5.3.1.3. *Refinement of the time-frequency pattern*

Nevertheless, the output of the filter/transduction module does not have the ideal characteristics of the representation which has been used for synthesis (see section 5.2.1): it tends to lack frequency and/or temporal resolution. The LIN network proposed by Shamma and mentioned above reinforces the spectral contrast [SHA 85]. Deng proposes the cross-correlation between neighboring channels in order to reinforce the formants' representation [DEN 88]. *Synchrony strands* by Cooke lead to a representation which is close to a sum of sine curves and which is well adapted for applying ASA principles (continuity of each strand, common fate, harmonicity, etc.) [COO 91]. These techniques can be interpreted as attempts to extract from the signal a representation close to the ideal representation which is used by psychoacousticians, according to the formulation of ASA's principles.

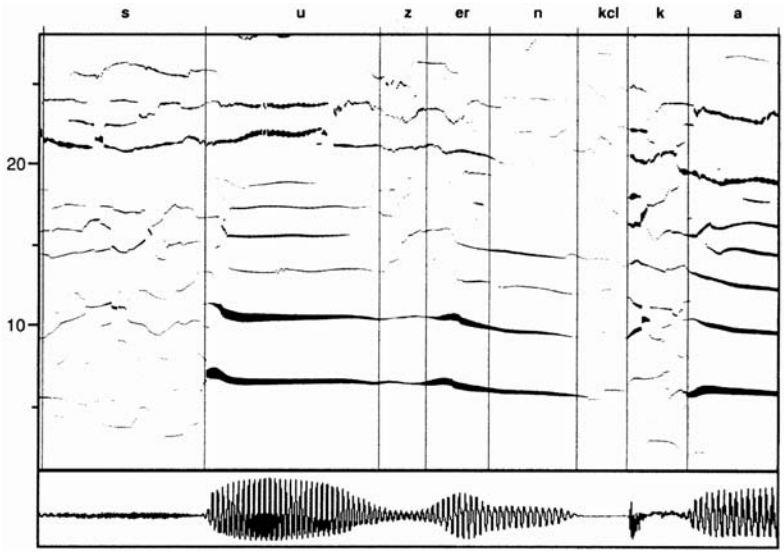


Figure 5.3. Synchrony strand *type representation, in response to a segment of speech (bottom). In the low frequencies, each strand corresponds to one harmonic, whereas in the high frequencies it corresponds to a formant (after [COO 91])*

5.3.1.4. *Additional dimensions*

If the time-domain smoothing is not too severe, the *temporal structure* of each channel stemming from the filter/transduction module can be exploited, leading to an enrichment of the representation with additional dimensions. Inspired by the binaural interaction model put forward by Jeffress, Lyon proposes to calculate the cross-correlation function between the channels from each ear [LYO 83]. Compared to a traditional time-frequency representation, this representation contains an additional dimension: the interaural delay. Maxima can appear at different positions along this dimension, corresponding to the azimuths of the various sources. Lyon samples the representation in terms of sections parallel to the frequency axis in order to isolate a particular source [LYO 83]. Similar attempts have been made since [BOD 96, PAT 96].

Another dimension comes into play if the *autocorrelation* function is calculated for each channel. This idea was originally proposed by Licklider in order to estimate the period in a perceptive pitch model [LIC 59]. In response to a periodic stimulus (as in voiced speech), maxima arise at locations corresponding to the period (and multiples of the period). This principle can be exploited in order to separate concurrent voice correlates. In response to several periodic stimuli (voices), some

channels may be dominated by one voice and others by another voice. By selecting channels according to the dominant periods, it is possible to isolate voices. Proposed by Weintraub [WEI 85], this idea was revisited by Mellinger [MEL 91], Meddis and Hewitt [MED 92], Brown [BRO 92a], Lea [LEA 92] and Ellis [ELL 96].

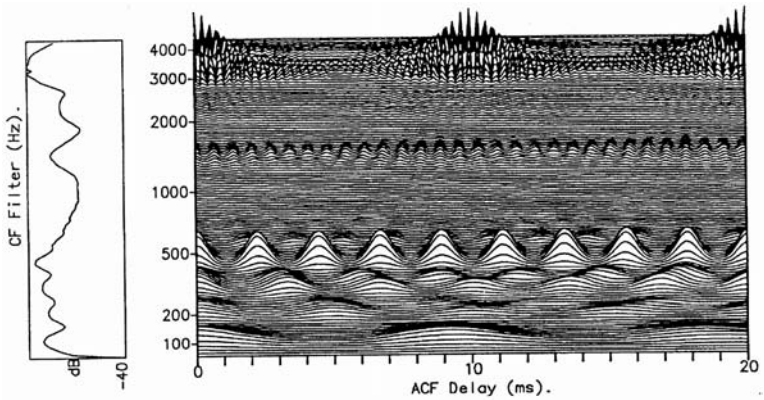


Figure 5.4. *Autocorrelation pattern corresponding to a mixture of vowels ([i] at 100 Hz and [o] at 112 Hz). Each line corresponds to one of the channels of the peripheral filter. Each channel is assigned to a vowel as a function of its dominant periodicity (after [LEA 92])*

Autocorrelation analyzes each channel with a sharp time-resolution, making it possible to resolve periodicity for the speech formants. However, the fine structure also reflects the resonance of the cochlear filters, which does not tell much about the signal. A *time-domain smoothing* operation enables the removal of the fine structure and (hopefully) retains only the modulations reflecting the fundamental period. These modulations can then be evaluated by autocorrelation or by other methods: zero-crossing [COO 91], Fourier transform [MEY 96, MEY 97]. The *modulation spectrum* of various parameters (physiological features, LPC coefficients, cepstral coefficients, etc), considered to be temporal sequences, has recently created a strong interest, in particular in the field of speech recognition [GRE 96, HER 94, KAN 98, NAD 97].

Among other transformations, let us mention the frequency transition map by Brown and the onset maps by Mellinger, Brown or Ellis, which aim at localizing abrupt time changes that may correspond to the beginning of a sound [BRO 92a, ELL 96, MEL 91].

Each additional dimension enriches the representation. If the acoustic pressure at one ear is dependent on *one* dimension (the time), the set of peripheral channels is dependent on *two* dimensions (time, frequency). When taken into account, the binaural correlation and the autocorrelation (or the modulation spectrum) lead to

four dimensions: time, frequency, interaural delay and modulation frequency. This "dimensional explosion" is motivated by the hope that cues for concurrent sounds will become separable if the dimensionality is high enough.

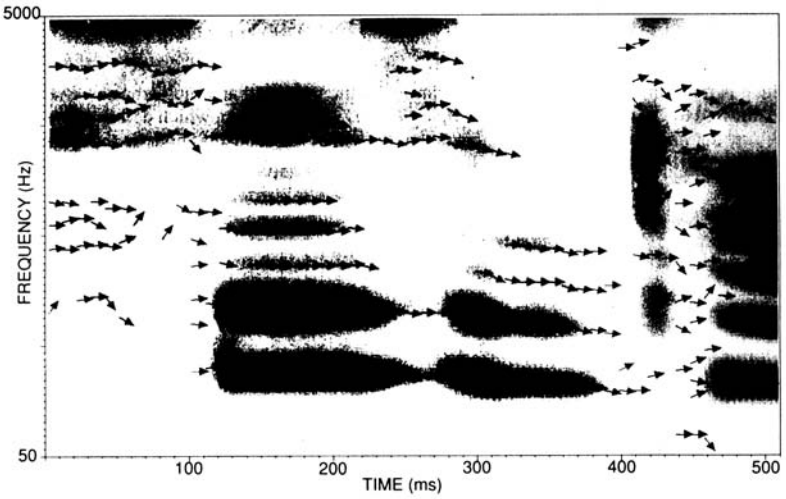


Figure 5.5. *Frequency transition map corresponding to a speech signal (same segment as in Figure 5.3). The arrows indicate the orientation estimated by a time-frequency orientation filter bank (after [BRO 93])*

5.3.1.5. *Basic abstractions*

Most CASA models start with rich, weakly constrained representations (see previous section), and then try to organize information as basic objects, following for example the ASA principles. Synchrony strands proposed by Cooke [COO 91] result from the application of a time continuity constraint to the components in the spectral representation. The principle of harmonicity grouping translates into *periodicity groups* in Cooke and Brown's approach [COO 92] or into *wefts* in Ellis' work [ELL 96]. The principle of attack synchrony is used by Brown to form auditory objects [BRO 92a].

5.3.1.6. *Higher-order organization*

The organization process carries on hierarchically, until the complete partition of the information into sources is obtained. Some models use a purely bottom-up (*data-driven*) process, while other models claim a more complex, top-down strategy calling for artificial intelligence techniques [ELL 96, GOD 97, KAS 97, NAK 97]. The drawback of complex strategies is twofold: they are opaque

and they tend to react catastrophically – in the sense that a small disruption of the system's input conditions may produce large changes of its state. However, they are essential for handling the whole set of information sources and hypotheses which take place in the organization of an auditory scene.

5.3.1.7. *Schemes*

Most CASA systems are *data-driven* and rely on ASA principles of the *primitive* type. *Top-down* approaches, relying on *scheme*-based ASA principles, are rare. It is worth mentioning Ellis' proposal to use a speech recognition system in order to guide auditory scene analysis [ELL 97]. When the speech component of the auditory scene is recognized, its contribution to the scene can be determined and the rest of the scene can be analyzed more accurately.

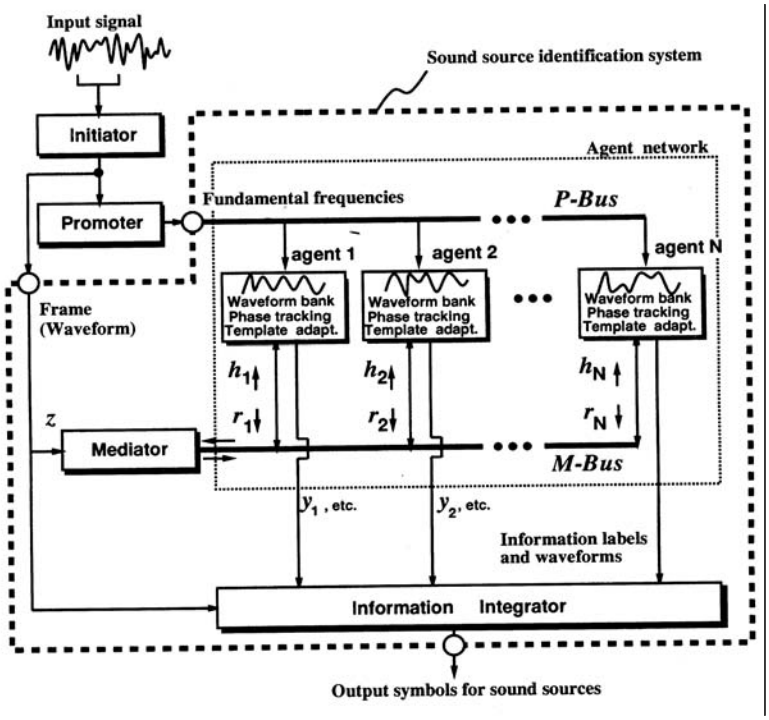
5.3.1.8. *The problem of shared components*

Whatever the richness and dimensionality of the representation, it may be that the assignment of a given element is ambiguous. Strategies vary depending on whether they assign this element to one source only (exclusive allocation), to both sources (duplex assignment) or to none at all. It is also possible to *split* the element according to some continuity criterion in the time or frequency domain [WEI 85]. Such a split can be seen as a failure of the representation, which has not succeeded in partitioning the acoustic information into atomic elements assignable to each source.

5.3.1.9. *The problem of missing components*

Theoretical reasons (that are unfortunately confirmed in practice) tell us that it is impossible to reach perfect separation in each and every situation. For instance, too close frequency components will be confused and assigned to one of the sources at the expense of another. Such masked or uncertain portions will be missing in the representation of the separated sources. Two approaches are possible to address this problem: 1) to re-create the missing information by interpolation or extrapolation from the acoustic or cognitive context [ELL 96, MAS 97]; or 2) to mark the corresponding portion as missing and to ignore it in the subsequent operations, for instance by assigning a zero weight in the pattern recognition step [COO 97, LIP 98, MOR 98].

The first approach, sometimes motivated by an over-literal interpretation of the notion of *phonemic restoration*, may be justified when re-synthesis is intended. The second approach is preferable in speech recognition applications.



Comment [JS1]: The figure is slightly skewed to the left. Would it be possible to provide a straighter version?

Figure 5.6. *An example of a CASA system structure: the Ipanema system for music analysis. Each agent is dedicated to the tracking of a particular aspect of the signal, under the control of a "mediator" (after [KAS 96])*

5.4. Critique of the CASA approach

The CASA approach is fertile, but it has weaknesses and pitfalls which need to be identified in order to be avoided.

5.4.1. *Limitations of ASA*

ASA is based on the concept that an auditory scene can be treated as a visual scene and that the Gestalt principles can be transposed from the vision domain, provided an adequate representation is chosen and a few adjustments are made to take into account the specificities of the acoustic domain. This idea can however lead to erroneous intuitions.

For example, let us consider harmonicity, which is a key grouping principle in ASA and is widely used in CASA models. ASA would state that spectral or temporal regularities of a harmonic target constitute a pattern which is easy to extract from the background (typically inharmonic or with a different fundamental frequency). Harmonicity would confer to a target some sort of texture which would facilitate its identification. This is not the case: many experiments show that the harmonicity of the *background* (or masking sound) does indeed facilitate segregation, but the harmonicity of the target has hardly any effect [DEC 95, DEC 97b, LEA 92, SUM 92c]. It can also been shown that the target's harmonicity is of limited usefulness in separating concurrent voices in a speech recognition task, and is less useful for the target than for the interference signal [DEC 93b, DEC 94].

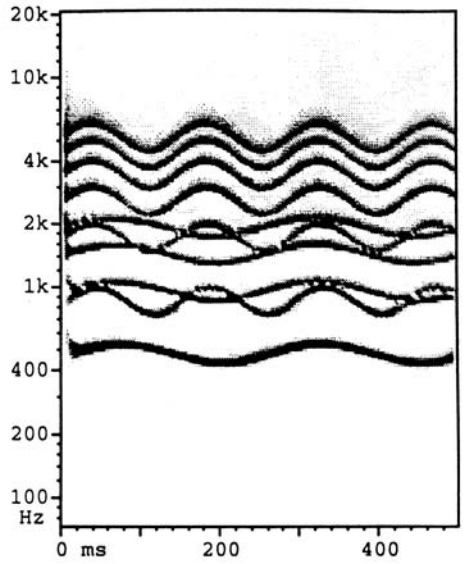


Figure 5.7. *The frequency modulation incoherence between two concurrent sources is exploited by the music analysis system (by [MEL 91]). Psychoacoustic experiments have shown however that this information is not used by the auditory system. This is an example where a CASA system exploits a Gestalt principle (common fate of the frequencies) that may not actually be used by the auditory system*

A second example is the Gestalt principle of *common fate*, which would mean that a spectrum made of components that vary in parallel (coherent frequency modulation) form a particular pattern which is especially easy to distinguish from a static background, or a background that would vary differently. Modulating a target which is not coherent with the background should thus facilitate its identification.

Once again, this is not the case: experiments show that frequency modulation has barely no other effect than the instantaneous F_0 differences induced by the modulation [CAR 94, DAR 95, DEM 90, MAR 97, MCA 89, SUM 92b].

One more example: the quality of a target's binaural correlation governs the accuracy with which it can be localized. We may think that this facilitates its segregation, whatever the nature of the background. Here again, this is not the case: segregation depends on the binaural correlation of the *masking sound* and not of the target. A well correlated masking sound is easy to cancel [COL 95, DUR 63]. Curiously, it is not necessary that the correlation be consistent across the various frequency channels [CUL 95].

5.4.2. *The conceptual limits of "separable representation"*

As mentioned above, the purpose of the initial, enriched representation is to allow correlates of different sources to be perceptually separated, by assigning elements of the representation to one source or another. However this goal is not always attained. Many authors have been confronted with the need to split elements of the representation, such as channels of a filterbank, and share them between sources [COO 91, ELL 96, PAR 76, WEI 85]. We may thus wonder whether the separable representation is in itself necessary. For example, the authors of [DEC 97b] have shown that Meddis and Hewitt's model [MED 92], that operates on a separable frequency-delay-time representation, could not explain all the F_0 difference effects on vowel segregation. Conversely, a model that operates on the time structure of nervous discharges in each frequency channel accounts well for segregation phenomena [DEC 97b]. This model performs better, but does not use a separable representation. Another example is the estimation of the fundamental frequencies of simultaneous sounds (for instance, the notes played by instruments which play together), which can be achieved without resorting to a separable representation of the time-frequency or autocorrelation type [DEC 93a, DEC 98].

Separable time-frequency-correlation representations, or the like, are often needed in CASA models. However, they are neither a panacea, nor a must, for auditory organization tasks.

5.4.3. *Neither a model, nor a method?*

The CASA approach offers a fertile and open field for new experiments, ideas, models and methods. This is not risk free. At best, the CASA specialist is well versed in the auditory sciences (psychoacoustics, physiology, etc.) and yet perfectly in gear with the application domain. At worst, he is neither. It is unfortunately

common to see an unrealistic model being defended in the name of "efficiency", or a poorly performing method on the basis that "this is the way the ear works".

Modeling approaches of auditory processes (whether computational or not) are flourishing and it is not always easy to understand the specificity of the CASA model. On the other hand, many techniques exist for source separation, noise reduction, etc. (in particular, *blind separation*), which are not affiliated to the CASA framework. They do not necessarily correspond to perceptual mechanisms, but this does not mean that they are less effective.

5.5. Perspectives

In spite of these weaknesses, the CASA approach contributes to the understanding of perceptual processes and to the design of new concepts in signal processing. Four recent developments are of particular interest.

5.5.1. *Missing feature theory*

There are situations when a CASA system (or similar) does not succeed in restoring some parts of the target signal. The corresponding data are missing. Their replacement by a zero value would alter the meaning of the pattern (for example, in a speech recognition system). Their replacement by an average over time or frequency is hardly a better solution. In some cases, interpolation or extrapolation from the context is justified. However, the optimal solution for a pattern recognition task consists in *ignoring* the missing data by assigning them a zero *weight* [AHM 93, COO 94, COO 96, COO 97, DEC 93b, GRE 96, LIPP 97, MOR 98].

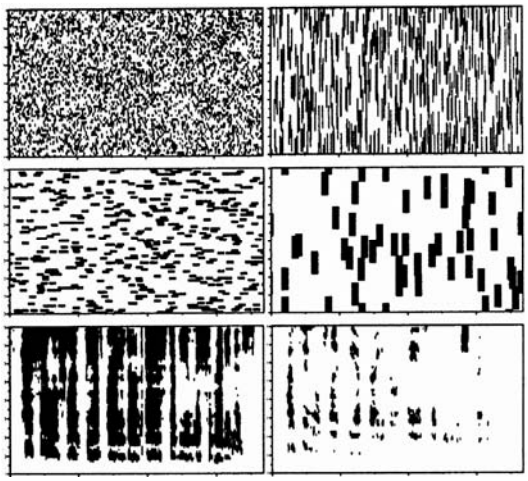


Figure 5.8. *Time-frequency masks used in missing data experiments. Areas in black correspond to available information, while the rest is missing. Recognition rates remain high even with a suppression rate of 80%. It is also possible to carry out training on incomplete data (after [COO 96])*

According to this approach, the CASA module delivers a *reliability map* to the recognition module. The latter module must be in the position to exploit it, which may raise difficulties in practice. For instance, many recognition systems use *cepstral* parameters, the advantage of which is to be orthogonally distributed and to allow the use of HMM (hidden Markov models) with diagonal covariance matrices. A reliability map in the *spectral* domain cannot be directly exploited by such a system. The use of spectral parameters instead of cepstral ones raises other problems [MOR 96].

The correct use of *missing feature* techniques is certainly a key to the practical application of the CASA approach. They can also be useful in a wider context, for instance for integrating information from different modalities. For example, an audiovisual recognition system can attribute a small weight to the image when the speaker's face is hidden, or a small weight to the audio modality when the speech sound is masked by noise.

5.5.2. *The cancellation principle*

Conventionally, ASA uses the structure of target sounds (for instance, their periodicity) in order to extract them from a non-structured (or differently-structured) environment. However, it has been observed that this approach is not particularly

effective, and that it is not the manner in which the auditory system proceeds. Let us consider the case of two microphones recording two sources with distinct azimuths. A system exploiting the position of the target will yield at best (by beam-forming) a 6 dB signal-to-noise ratio reduction, whereas a system using the knowledge of the position of the interference can theoretically reach an infinite signal-to-noise ratio (even though, in practice, the improvement is limited in the case of reverberation or multiple maskers). Similarly, a system exploiting the target periodicity for reinforcing it will not perform as efficiently as a system exploiting the periodicity of the background to cancel it [DEC 93a, DEC 93b]. The auditory system exploits the background's periodicity rather than that of the target [DEC 95, DEC 97c, LEA 92, SUM 92c]. The cancellation criterion is close to that used by blind separation techniques. Scene analysis by successive cancellation steps is a characteristic of Nakatani's system [NAK 95a, NAK 95b, NAK 97].

Cancellation offers in some cases an infinite rejection (i.e. an infinite improvement of the target/background ratio), but it usually introduces some distortion of the target. For example, the shared (or masked) components are suppressed. Missing feature techniques are useful in this case.

5.5.3. *Multimodal integration*

The development of multimodal speech recognition opens the way for multimodal scene analysis, which would be more than the simple juxtaposition of visual and auditory analysis modules [OKU 99a]. There again, missing feature approaches are promising for the integration of modal data with variable reliability.

5.5.4. *Auditory scene synthesis: transparency measure*

ASA can also be approached from a radically different angle, i.e. from the viewpoint of an audio scene designer. When the audio material is assembled and mixed, it may happen that one of the ingredients is particularly dominant and has a strong masking effect that makes the auditory scene rather confused. Taking into account the various parameters revealed by ASA enables the prediction of the masking power of a source according to its physical characteristics. An *audio transparency* measure would be useful for the designer, to help in choosing the optimal ingredients. Such a measure has been proposed in the future MPEG7 standard for the description of multimedia data [DEC 99b].

5.6. Bibliography

[AHM 93] AHMAD S., TRESP V., "Some solutions to the missing feature problem in vision", in S.J. Hanson, J.D. Cowan and C.L. Giles (eds.), *Advances in Neural Information Processing Systems 5*, p. 393-400, San Mateo, Morgan Kaufmann, 1993.

[ASS 90] ASSMANN P.F., SUMMERFIELD Q., "Modeling the perception of concurrent vowels: vowels with different fundamental frequencies", *J. Acoust. Soc. Am.*, 88, p. 680-697, 1990.

[BER 95] BERTHOMMIER F., MEYER G., "Source separation by a functional model of amplitude demodulation", *Proc. of ESCA Eurospeech*, p. 135-138, 1995.

[BOD 96] BODDEN M., RATEIKSHEK K., "Noise-robust speech recognition based on a binaural auditory model", *Proc. of Workshop on the auditory basis of speech perception*, Keele, p. 291-296, 1996.

[BRE 90] BREGMAN A.S., *Auditory scene analysis*, MIT Press, Cambridge, 1990.

[BRO 82] BROKX J.P.L., Nooteboom S.G., "Intonation and the perceptual separation of simultaneous voices", *Journal of Phonetics*, 10, p. 23-36, 1982.

[BRO 92a] BROWN G.J., Computational auditory scene analysis: a representational approach, Sheffield, Department of Computer Science, unpublished PhD Thesis, 1992.

[BRO 92b] BROWN G.J., Cooke M.P., "Computational auditory scene analysis: grouping sound sources using common pitch contours", *Proc. Inst. of Acoust.*, 14, p. 439-446, 1992.

[BRO 93] BROWN G.J., COOKE M., "Physiologically-motivated signal representations for computational auditory scene analysis", in M. Cooke, S. Beet and M. Crawford (eds.), *Visual Representations of Speech Signals*, Chichester, John Wiley and Sons, p. 181-188, 1993.

[CAR 94] CARLYON R., "Further evidence against an across-frequency mechanism specific to the detection of frequency modulation (FM) incoherence between resolved frequency components", *J. Acoust. Soc. Am.*, 95, p. 949-961, 1994.

[CHE 53] CHERRY E.C., "Some experiments on the recognition of speech with one, and with two ears", *J. Acoust. Soc. Am.*, 25, p. 975-979, 1953.

[COL 95] COLBURN H.S., "Computational models of binaural processing", in H. Hawkins, T. McMullin, A.N. Popper and R.R. Fay (eds.), *Auditory Computation*, New York, Springer-Verlag, p. 332-400, 1995.

[COO 91] COOKE M.P., Modeling auditory processing and organization, Sheffield, Department of Computer Science, unpublished Thesis, 1991.

[COO 93] COOKE M.P., BROWN G.J., "Computational auditory scene analysis: exploiting principles of perceived continuity", *Speech Comm.*, 13, p. 391-399, 1993.

[COO 94] COOKE M., GREEN P., ANDERSON C., ABBERLEY D., Recognition of occluded speech by hidden markov models, University of Sheffield Department of Computer Science, Technical report, TR-94-05-01, 1994.

[COO 96] COOKE M., MORRIS A., GREEN P., "Recognising occluded speech", *Proc. Workshop on the Auditory Basis of Speech Perception*, Keele, p. 297-300, 1996.

[COO 97] COOKE M., MORRIS A., GREEN P., "Missing data techniques for robust speech recognition", *Proc. ICASSP*, p. 863-866, 1997.

[COO 99] COOKE M., ELLIS D.P.W., "The auditory organization of speech and other sources in listeners and computational models", *Speech Communication*, 35, p. 141-177, 2001.

[CUL 95] CULLING, J.F., SUMMERfiELD Q., "Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay", *J. Acoust. Soc. Am.*, 98, p. 785-797, 1995.

[DAR 95] DARWIN C.J., CARLYON R.P., "Auditory grouping", in B.C.J. Moore (ed.), *Handbook of perception and cognition: Hearing*, New York, Academic Press, p. 387-424, 1995.

[DEC 93a] DE CHEVEIGNÉ A., "Separation of concurrent harmonic sounds: fundamental frequency estimation and a time-domain cancellation model of auditory processing", *J. Acoust. Soc. Am.*, 93, p. 3271-3290, 1993.

[DEC 93b] DE CHEVEIGNÉ A., Time-domain comb filtering for speech separation, ATR Human Information Processing Laboratories, Technical Report, TR-H-016, 1993.

[DEC 94] DE CHEVEIGNÉ A., KAWAHARA H., AIKAWA K., LEA A., "Speech separation for speech recognition", *Journal de Physique*, IV 4, C5-545-C5-548, 1994.

[DEC 95] DE CHEVEIGNÉ A., MCADAMS S., LAROCHE J., ROSENBERG M., "Identification of concurrent harmonic and inharmonic vowels: a test of the theory of harmonic cancellation and enhancement", *J. Acoust. Soc. Am.*, 97, p. 3736-3748, 1995.

[DEC 97a] DE CHEVEIGNÉ A., "Concurrent vowel identification III: A neural model of harmonic interference cancellation", *J. Acoust. Soc. Am.*, 101, p. 2857-2865, 1997.

[DEC 97b] DE CHEVEIGNÉ A., KAWAHARA H., TSUZAKI M., AIKAWA K., "Concurrent vowel identification I: effects of relative level and F0 difference", *J. Acoust. Soc. Am.*, 101, p. 2839-2847, 1997.

[DEC 97c] DE CHEVEIGNÉ A., MCADAMS S., MARIN C., "Concurrent vowel identification II: effects of phase, harmonicity and task", *J. Acoust. Soc. Am.*, 101, p. 2848-2856, 1997.

[DEC 98] DE CHEVEIGNÉ A., "Cancellation model of pitch perception", *J. Acoust. Soc. Am.*, 103, p. 1261-1271, 1998.

[DEC 99a] DE CHEVEIGNÉ A., KAWAHARA H., "Multiple period estimation and pitch perception model", *Speech Communication*, 27, p. 175-185, 1999.

[DEC 99b] DE CHEVEIGNÉ A., SMITH B., A "sound transparency" descriptor ISO/IEC JTC1/SC29/WG11, MPEG99/m5199, 1999.

[DUR 63] DURLACH, N.I., "Equalization and cancelation theory of binaural masking-level differences", *J. Acoust. Soc. Am.*, 35, p. 1206-1218, 1963.

[ELL 96] ELLIS D., Prediction-driven computational auditory scene analysis, MIT, unpublished Thesis, 1996.

[ELL 97] ELLIS D.P.W., "Computational auditory scene analysis exploiting speech-recognition knowledge", *Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio*, Mohonk, 1997.

[GRE 95] GREEN P.D., COOKE M.P., CRAWFORD M.D., "Auditory scene analysis and hidden Markov model recognition of speech in noise", *Proc. IEEE-ICASSP*, p. 401-404, 1995.

[GRE 97] GREENBERG, "Understanding speech understanding: towards a unified theory of speech perception", *Proc. ESCA Workshop on the auditory basis of speech perception*, Keele, p. 1-8, 1997.

[HAR 96] HARTMANN W.M., "Pitch, periodicity, and auditory organization", *J. Acoust. Soc. Am.*, 100, p. 3491-3502, 1996.

[HEL 77] HELMHOLTZ H. V., (1877). *On the sensations of tone (English translation A.J. Ellis, 1954)*, New York, Dover.

[HER 94] HERMANSKY H., MORGAN N., "RASTA processing of speech", *IEEE trans Speech and Audio Process.*, 2, p. 578-589, 1994.

[HOL 88] HOLDSWORTH J., NIMMO-SMITH I., PATTERSON R.D., RICE P., Implementing a gammatone filter bank, MRC Applied Psychology Unit technical report, SVOS Final Report, Appendix C, 1998.

[HOL 90] HOLDSWORTH J., Two dimensional adaptive thresholding, APU AAM-HAP report, Technical Report, vol. 1, Appendix 4, 1990.

[HOL 92] HOLDSWORTH J., SCHWARTZ J.-L., BERTHOMMIER F., PATTERSON R.D., "A multirepresentation model for auditory processing of sounds", in Y. Cazals, L. Demany and K. Horner (eds.), *Auditory physiology and perception*, Oxford, Pergamon Press, p. 447-453, 1992.

[JOR 98] JORIS P.X., YIN T.C.T., "Envelope coding in the lateral superior olive. III. comparison with afferent pathways", *J. Neurophysiol.*, 79, p. 253-269, 1998.

[KAN 98] KANADERA N., HERMANSKY H., ARAI T., "On properties of the modulation spectrum for robust automatic speech recognition", *Proc. IEEE-ICASSP*, p. 613-616, 1998.

[LEA 92] LEA A., Auditory models of vowel perception, Nottingham University, unpublished Thesis, 1992.

[LIC 59] LICKLIDER J.C.R., "Three auditory theories", S. Koch (ed.), *Psychology, a study of a science*, New York, McGraw-Hill, I, p. 41-144, 1959.

[LIP 97] LIPPMANN R.P., CARLSON B.A., "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise", *Proc. ESCA Eurospeech*, KN-37-40, 1997.

[LYO 83] LYON, R.F., "A computational model of binaural localization and separation", W. Richards (ed.), *Natural computation*, Cambridge, Mass, MIT Press, p. 319-327, 1983.

[LYO 84] LYON R., "Computational models of neural auditory processing", *Proc. IEEE ICASSP*, 36.1.(1-4), 1984.

[LYO 91] LYON R., "Automatic gain control in cochlear mechanics", in P. Dallos, C.D. Geisler, J.W. Mathews (eds.), *Mechanics and Biophysics of Hearing*, M.A. Ruggero and C.R. Steele, New York, Springer-Verlag, 1991.

[MAR 82] MARR D., "Representing and computing visual information", in P.H. Winston and R.H. Brown, *Artificial Intelligence: an MIT perspective*, Cambridge, Mass, MIT Press, 2, p. 17-82, 1982.

[MAR 97] MARIN C., DE CHEVEIGNÉ A., "Rôle de la modulation de fréquence dans la séparation de voyelles", *Proc. Congrès Français d'Acoustique*, p. 527-530, 1997.

[MCA 84] MCADAMS S., Spectral fusion, spectral parsing, and the formation of auditory images, Stanford University, unpublished Thesis, 1984.

[MCA 89] MCADAMS S., "Segregation of concurrent sounds I: effects of frequency modulation coherence", *J. Acoust. Soc. Am.*, 86, p. 2148-2159, 1989.

[MED 88] MEDDIS R., "Simulation of auditory-neural transduction: further studies", *J. Acoust. Soc. Am.*, 83, p. 1056-1063, 1988.

[MED 92] MEDDIS R., HEWITT M.J., "Modeling the identification of concurrent vowels with different fundamental frequencies", *J. Acoust. Soc. Am.*, 91, p. 233-245, 1992.

[MEL 91] MELLINGER D.K., Event formation and separation in musical sound, Stanford Center for Computer Research in Music and Acoustics, unpublished Thesis, 1991.

[MEY 96] MEYER G., BERTHOMMIER F., "Vowel segregation with amplitude modulation maps: a re-evaluation of place and place-time models", *Proc. ESCA Workshop on the Auditory Basis of Speech Perception*, Keele, p. 212-215, 1996.

[MEY 97] MEYER G.F., PLANTE F., BERTHOMMIER F., "Segregation of concurrent speech with the reassigned spectrum", *Proc. IEEE ICASSP*, p. 1203-1206, 1997.

[MOR 98] MORRIS A.C., COOKE M.P., GREEN P.D., "Some solutions to the missing feature problem in data classification, with application to noise robust ASR", *Proc. ICASSP*, p. 737-740, 1998.

[NAD 97] NADEU C., PACHÈS-LEAL P., JUANG B.-H., "Filtering the time sequences of spectral parameters for speech recognition", *Speech Comm.*, 22, p. 315-332, 1997.

[NAK 95a] NAKATANI T., OKUNO H.G., KAWABATA T., "Residue-driven architecture for computational auditory scene analysis", *Proc. IJCAI*, p. 165-172, 1995.

[NAK 95b] NAKATANI T., GOTO M., ITO T., OKUNO H.G., "Multi-agent based binaural sound stream segregation", *Proc. IJCAI Workshop on Computational Auditory Scene Analysis*, p. 84-91, 1995.

[NAK 96] NAKATANI T., GOTO M., OKUNO H. G., "Localization by harmonic structure and its application to harmonic stream segregation", *Proc. IEEE ICASSP*, p. 653-656, 1996.

[NAK 97] NAKATANI T., KASHINO K., OKUNO J.G., "Integration of speech stream and music stream segregations based on a sound ontology", *Proc. IJCAI Workshop on Computational Auditory Scene Analysis*, Nagoya, p. 25-32, 1997.

[OKU 99a] OKUNO H.G., NAKAGAWA Y., KITANO H., "Incorporating visual information into sound source separation", *Proc. International Workshop on Computational Auditory Scene Analysis*, 1999.

[OKU 99b] OKUNO H.G., IKEDA S., NAKATANI T., "Combining independant component analysis and sound stream segregation", *Proc. International Workshop on Computational Auditory Scene Analysis*, 1999.

[PAR 76] PARSONS, T.W., "Separation of speech from interfering speech by means of harmonic selection", *J. Acoust. Soc. Am.*, 60, p. 911-918, 1976.

[PAT 92] PATTERSON R.D., ROBINSON K., HOLDSWORTH J., MCKEOWN D., ZHANG C., ALLERHAND M., "Complex sounds and auditory images", in Y. Cazals, K. Horner and L. Demany (eds.), *Auditory Physiology and Perception*, Oxford, Pergamon Press, p. 429-446, 1992.

[PAT 96] PATTERSON R., ANDERSON T.R., FRANCIS K., "Binaural auditory images and a noiseresistant, binaural auditory spectrogram for speech recognition", *Proc. Workshop on the Auditory Basis of Speech Perception*, Keele, p. 245-252, 1996.

[ROS 97] ROSENTHAL D.F., OKUNO H.G., *Computational Auditory Scene Analysis*, Lawrence Erlbaum, 1997.

[SCH 83] SCHEFFERS M.T.M., Sifting vowels, Gröningen University, Thesis, 1983.

[SEN 85] SENEFF S., Pitch and spectral analysis of speech based on an auditory synchrony model, MIT, unpublished Thesis (Technical Report 504), 1985.

[SHA 85] SHAMMA S.A., "Speech processing in the auditory system I: he representation of speech sounds in the responses of the auditory nerve", *J. Acoust. Soc. Am.*, 78, p. 1612-1621, 1985.

[SLA 93] SLANEY M., An efficient implementation of the Patterson-Holdsworth auditory filter bank, Apple Computer Technical Report, 35, 1993.

[SLA 95] SLANEY, M., "A critique of pure audition", *Proc. Computational Auditory Scene Analysis Workshop*, IJCAI, Montreal, 1995.

[SUM 90] SUMMERfiELD Q., LEA A., MARSHALL D., "Modelling auditory scene analysis: strategies for source segregation using autocorrelograms", *Proc. Institute of Acoustics*, 12, p. 507-514, 1990.

[SUM 92a] SUMMERfiELD Q., CULLING J.F., "Auditory segregation of competing voices: absence of effects of FM or AM coherence", *Phil. Trans. R. Soc. Lond.*, B 336, p. 357-366, 1992.

[SUM 92b] SUMMERfiELD Q., "Roles of harmonicity and coherent frequency modulation in auditory grouping", in M.E.H. Schouten (ed.), *The Auditory Processing of Speech: From Sounds to Words*, Berlin, Mouton de Gruyter, p. 157-166, 1992.

[SUM 92c] SUMMERfiELD Q., CULLING J.F., "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency", *Proc. 124th Meeting of the ASA*, 2317(A), 1992.

[WAN 95] WANG A.L.-C., Instantaneous and frequency-warped signal processing techniques for auditory source separation, CCRMA (Stanford University), unpublished Thesis, 1995.

[WAR 70] WARREN R.M., "Perceptual restoration of missing speech sounds", *Science*, 167, p. 392-393, 1970.

[WAR 72] WARREN R.M., OBUSEK C.J., ACKROFF J.M., "Auditory induction: perceptual synthesis of absent sounds", Science, 176, p. 1149-1151, 1972.

[WEI 85] WEINTRAUB M., A theory and computational model of auditory monaural sound separation, Stanford University, unpublished Thesis, 1985.

[YOS 96] YOST W.A., DYE R.H., SHEFT S., "A simulated 'cocktail party' with up to three sound sources", *Perception and Psychophysics*, 58, p. 1026-1036, 1996.