The auditory system as a "separation machine"

A. de Cheveigné

CNRS-Ircam, 1 place Igor Stravinsky,
75004, Paris, France.
cheveign@ircam.fr

1. Introduction

This paper explores the hypothesis that the auditory system is designed to *separate* sounds rather than just detect, discriminate, or recognize them.

Szentagothai and Arbib (1975) describe a primitive fish-like organism with a very simple nervous system. Sensors placed on either side of the head are connected to fins on the opposite side by a neuron. When food is sensed on one side a signal is transmitted to the opposite fin, the fish turns towards the food, and this orientation is maintained by the balance of bilateral activation until the food is reached. For such a simple organism, perception and action are equivalent, and Szentagothai and Arbib suggest that the same is basically true for higher organisms, with additional levels of inhibition that complexify behavior. The crossed pathway between brain and body would be a heritage of this primitive structure.

This simple organism can survive in a world where things visible are also edible. If the world contains predators in addition to prey, a more complex behavior is required. Based on what it sees, the organism must decide to activate the contra-lateral fin to get closer and eat, or the ipsilateral fin to escape from being eaten. This more sophisticated behavior requires *discrimination* between sensible objects. If the inventory of objects and actions were large, one might also speak of recognition or identification.

Detecting organisms and discriminating organisms both require that prey and/or predators appear in isolation. If both predator and prey (or several of each) appear together the organism won't know how to react. To survive in a densely populated world the organism must be capable of *segregation*. Segregation is the ability to selectively process sensory evidence by parts, assigning each part to a "source" within a model of the world. Segregation allows a cat to hear the faint sounds made by a mouse in the rustling grass, and it might be of use to the mouse in that same situation.

For detection and discrimination, all sensory evidence is attributed to one source. For segregation, it is *partitioned* and shared between sources. The partition must precede extraction of source qualities, yet it also appears to depend upon those qualities, a paradox emphasized by Bregman (1990). Classic psychoacoustics has concentrated on detection and discrimination, and only recently has segregation come to the forefront with the ideas of Bregman and others. A century earlier Helmholtz (1877) had asked how one hears the quality of an instrument playing among others, but that question was put aside for the following century. In summary, segregation is essential for survival but harder to account for than detection or discrimination. The hypothesis explored here is that the auditory system is in large part designed for this task. An element of the argument is *missing-feature theory*.

2. Missing-feature theory

"Missing-feature theory" aims to allow an artificial speech recognition system to adapt to the incomplete evidence provided by a computational auditory scene analysis (CASA) front-end (Cooke et al., 1996; Lippmann, 1997; Morris et al., 1998). If speech is corrupted by interference, the CASA front-end may successfully suppress the interference but parts of the speech are likely to be suppressed at the same time, so the recognizer will behave poorly. Three options are available, that can be qualified as "bad", "better", or "optimal". A "bad" option is to set the missing values to zero or to an arbitrary constant. A "better" option is to perform some form of interpolation or extrapolation from neighboring, intact data. This might be the best course if the aim is to resynthesize speech after segregation. However the "optimal" option for recognition is arguably to *ignore* missing parts. Interpolation is essentially a principled guess, and as such it may be wrong. Ignoring missing data is a safer course.

Missing-feature theory has practical applications in speech recognition and vision (Ahmad and Tresp, 1993), but it is of wider use within perception models. For example it may be used to explain the continuity illusion and phonemic "restoration" effects without the need to postulate perceptual synthesis of low-level correlates. A similar principle can be applied to cross-modal integration of information, each modality being weighted according to its reliability (Massaro, 1990). Missing-feature theory is useful in the context of segregation to handle the incomplete patterns retrieved by a segregation mechanism.

3. Tonotopy in the auditory system

This section and the next review physiological evidence for the two segregation principles considered in this paper, "channel selection" and "channel splitting".

The orderly distribution of characteristic frequencies (CF) within the cochlea (tonotopy) is reflected at many levels of the auditory system. The three major divisions of the cochlear nucleus (AVCN, PVCN, DCN) are tonotopically organized, as are nuclei of the superior olivary complex (MOC, LOC, MNTB), the dorsal nucleus of the lateral lemniscus (DNLL), the central nucleus of the inferior colliculus (ICC) and, at least in anesthetized animals, the ventral nucleus of the medial geniculate body (vMGB) and several fields of cortex, particularly the primary auditory field (AI). Efferent pathways are also tonotopically organized, in particular the medial and lateral olivocochlear pathways that project, respectively, to the outer hair cells and inner hair cell afferents (Cant, 1992; Helfert and Aschoff, 1997; Rouiller, 1992; de Ribaupierre, 1997; Clarey et al., 1992).

Traditionally, cochlear analysis is assimilated to a Fourier transform, and tonotopically organized nuclei are assumed to repeat copies of a spectral representation. Tonotopy is taken as evidence for the importance of a spectral code. However it is not clear why such a code must be repeated at every level. An alternative hypothesis, explored in this paper, is that peripheral analysis splits the incoming sound into an array of partly redundant band-limited channels which may be differentially weighted according to their relevance or reliability (Hermansky, 1998).

4. Time in the auditory system

Auditory-nerve fibers synchronize to the fine structure of stimuli. Measures of synchrony tend to drop beyond 1-2 kHz, but they remain significant up to 4-6 kHz in mammals or 9 kHz in the barn owl. This upper frequency limit of synchrony does not necessarily determine the limit of temporal resolution: onset latencies of some cells of the cochlear nucleus (CN) have less than 100 μs standard deviation, and psychophysical experiments show that ITDs as small as 6 μs can be exploited (Irvine, 1992).

Some neural hardware seems to be designed for coding temporal information: specialized synapses, large cell bodies, and membranes with fast recovery. In CN, spherical bushy cells (SBC) are fed by single auditory-nerve fibers via the "end-bulbs of Held" that ensure secure transmission of every incoming spike with little loss of time resolution. Also in CN, globular bushy cells (GBC) are fed by small numbers of auditory-nerve fibers via similar secure synapses. Principal cells in the medial nucleus of the trapezoid body (MNTB) are fed via "calyces of Held" by thick myelinized fibers from GBCs in contralateral CN. In addition to these cells that faithfully relay the temporal structure of auditory-nerve activity, others, such as octopus cells in CN, enhance certain aspects of synchrony at the expense of others, and in particular respond to onsets with high temporal resolution (Schwartz, 1992; Joris and Yin, 1998; Oertel, 1999; Trussel, 1999; Sabatini, 1999).

SBCs and GBCs project from cochlear nucleus to many relays: ipsilateral and contralateral CN, the superior olivary complex (MNTB, LSO, MSO and periolivary nuclei), nuclei of the lateral lemniscus, and the inferior colliculus (IC). The inhibitory relay cells of MNTB project to LSO, MSO, VNLL and various periolivary nuclei (Schwartz, 1992; Romand and Avan, 1997, Helfert and Aschoff, 1997). High-resolution temporal patterns have been found at many levels up to IC. Logically, they should also be present in axonal projections from these levels, and time-domain signal processing may occur at the sites where such projections interact.

MSO is implicated in the time-domain processing of interaural time differences (ITDs) (Jeffress, 1948; Yin and Chan, 1990), and LSO is traditionally assigned the processing of interaural level differences (ILDs). The time-specialized circuits that feed LSO from ipsilateral CN and contralateral MNTB are hard to justify for static ILDs. However recent studies have suggested that LSO might play a role in processing *dynamic* ILDs (onsets) or processing of "multiplexed" evidence of concurrent sources (Joris and Yin, 1998). According to the latter suggestion, processing in LSO might embody the Equalization-Cancellation (EC) model of binaural unmasking of Durlach (1963).

There is little convergence between frequency channels in these circuits, as far as major excitatory inputs are concerned. SBCs are fed by single AN fibers, GBCs by small groups of presumably similar AN fibers, and MNTB principal cells by single GBC axons. This allows a degree of independence between the processing that occurs within different frequency channels. It is of interest to note that Culling and Summerfield (1995) have recently proposed a modified version of Durlach's EC model in which processing occurs within individual channels, based on criteria local to that channel. The model successfully explains a wide variety of binaural phenomena (Culling et al., 1998).

Processing binaural information is the role most commonly invoked for the time-specialized neural circuitry of CN/MNTB/LSO/MSO. This heavy investment is hard to justify for a function that is undeveloped in many animals. LSO, for example, is little developed in humans (Heffner and Heffner, 1992). As noted earlier, temporally specialized circuits have many projections other than MSO and LSO. It is unlikely that they are *all* involved in binaural processing. An alternative hypothesis is that time-domain segregation processes complement the across-channel segregation supported by tonotopy. The next two sections provide examples of monaural and binaural segregation models based on tonotopy and time.

5. Models of binaural segregation

Subjectively, it seems easier to attend sources that are spatially separated than sources coming from same spot. Binaural cues contribute to the "cocktail party effect" according to Cherry (1953), and binaural unmasking effects have been studied intensively in psychoacoustic experiments (Durlach, 1978). Binaural segregation models can be divided into two classes: channel-selecting and channel-splitting.

Channel-selection follows the ideas of Lyon (1983), themselves based on the Jeffress (1948) localization model. In Jeffress's model, an array of cross-correlation functions is formed, one for each peripheral channel. In response to an isolated source, a ridge appears in the array at a position that signals the azimuth of the source. Lyon applied the model to several sources with different azimuths. Supposing the sources have different spectral envelopes, the ridge appears in different positions in different channels. This information is used to *label* the channels and assign them to a source. The channel-selection principle has been used repeatedly in binaural models (for example Patterson et al., 1996). Channel-selection works hand in hand with peripheral analysis, and depends on it for actual segregation: features that are not resolved in the cochlea cannot be segregated by channel selection.

Channel-splitting is exemplified by the Equalization-Cancellation (EC) model of Durlach (1963), in which signals from both ears are equalized by scaling and delaying one relative to the other, and then subtracted. The remainder is used as a signal. Processing is presumably applied uniformly within every channel (this was not stated explicitly because the model was aimed at narrow-band phenomena). To the extent that filtering and EC operations are linear, they can conceptually be swapped, as if the EC operations were performed directly on the signal. Peripheral selectivity thus plays no major role in the original EC model. On the other hand, in the *modified EC model* of Culling and Summerfield (1995) equalization is performed independently within each channel based on channel-specific criteria. Peripheral selectivity has a role to play in this case.

A physiological implementation of the EC or modified EC models would involve time-domain interaction of neural signals with high temporal resolution. The result might be smoothed and treated as a spectral pattern (residual activity vs channel). Alternatively, the fine temporal structure of the output might be conserved and submitted to additional time-domain processing.

6. Models of harmonic segregation

A sound that is periodic (in time) or equivalently harmonic (in frequency) usually evokes a pitch sensation. Harmonicity is also exploited in the "cocktail party effect" to segregate voices and improve the intelligibility of speech in the presence of interference. When two harmonic sounds (two voices) compete, there are potentially two harmonic series to exploit. It turns out that the intelligibility of a voice depends mainly on the harmonic structure of the *competing* voice (Summerfield and Culling, 1992; Lea, 1992; de Cheveigné et al., 1997a,b). In other words, the harmonic structure of interference is exploited to suppress it, but the harmonic structure of a target does not help to enhance that target. Like binaural models, harmonic segregation models can be divided into two classes: channel-selection and channel-splitting.

Channel-selection, based on the ideas of Weintraub (1985), has been more recently developed by Meddis and Hewitt (1992). An array of autocorrelation functions (ACF) is calculated, one for each channel. The position of the 'period peak" of the ACF within a channel indicates the period that dominates it, and this allows the channel to be labeled as belonging to one source or the other.

Channel-splitting is performed in the concurrent vowel identification model of de Cheveigné (1993, 1997). Each channel is processed by a "neural cancellation filter" tuned to suppress the period of the interference. As for binaural cancellation models, the output of a tonotopic array of cancellation filters can be taken either as a spectral pattern (static or slowly varying), or as an array of time-domain patterns.

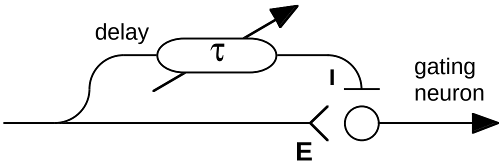

Fig. 1. Neural cancellation filter. The neuron is fed via two pathways, one direct and excitatory and the other delayed and inhibitory. Every spike arriving along the direct pathway is transmitted unless a spike arrives simultaneously along the delayed pathway. The filter reduces the probability of all-order intervals equal to the delay, and thus suppresses correlates of a source with that period.

7. Select or split?

Channel selection models can recover a source that dominates part of the spectrum. A source dominated by interference within *all* peripheral channels cannot be recovered, and channel selection models predict no segregation in that case. In a concurrent vowel identification experiment, we mixed synthetic vowels with amplitudes such that the spectral envelope of the stronger vowel dominated that of the weaker vowel over all the spectrum (de Cheveigné et al., 1997a). Identification of the weaker vowel was nevertheless greatly improved by a difference in fundamental frequency between vowels. A channel-splitting model is necessary to explain such results.

Supposing that channel selection and splitting mechanisms are available to the auditory system, source segregation is likely to involve both. The former are limited by peripheral selectivity, the latter by the dynamic range of time-domain neural processing. Peripheral filtering may for example fail to resolve components of the weaker source, but nevertheless raise the signal-to-noise ratio sufficiently, within certain channels, to allow a channel-splitting mechanism to work.

Channel-selection may recover parts of a target pattern, but other parts will be missing. Pattern matching may nevertheless proceed reliably if missing features are *ignored* within the pattern matching mechanism. Ignoring part of a pattern is not the same as setting it to zero. Likewise, channel-splitting mechanisms distort the patterns that they recover (this distortion is analogous to that caused by comb-filtering). To the extent that distortion is known, it may be taken into account by pattern-matching.

8. Summary and conclusion

In this paper the structure and properties of the auditory system were interpreted as serving the purpose of *segregating* sources. This interpretation is not exclusive of other roles, and may even lead to fruitful insights concerning those roles. For example, the cancellation filter of Fig. 1 developed for F0-guided segregation turns out to be effective when applied to pitch estimation (de Cheveigné, 1998).

Auditory Scene Analysis theory puts the emphasis on *grouping* elements homologous to partials or Fourier components. The availability of such elements is taken for granted, and the focus is put on finding criteria (for example drawn from Gestalt theory) to assemble them. In contrast, this paper supposes that the auditory system must work hard to divide representations that are essentially unitary into entities that faithfully represent individual sources. In a good model of the world, each important object of the world should be given its own perceptual correlates. The thesis of this paper is that the auditory system is designed to perform this difficult task. This goes beyond the classic view of an auditory system as a mere "estimator" of auditory qualities, or "recognizer" of patterns, and it offers other roles for cochlear selectivity than mere Fourier Analysis.

9. Acknowledgements

A preliminary version of this paper was presented at the ATR-EATS workshop supported by ATR Human Information Processing Research Laboratories (Japan).

10. References

Ahmad, S., and Tresp, V. (1993). Some solutions to the missing feature problem in vision, In *Advances in Neural Information Processing Systems 5*. Edited by S. J. Hanson, J. D. Cowan and C. L. Giles, San Mateo, Morgan Kaufmann, 393-400.

Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge, Mass., MIT Press.

Cant, N. B. (1992). The cochlear nucleus: neuronal types and their synaptic organization. In: D. B. Webster, A. N. Popper and R. R. Fay (Eds.), *The mammalian auditory pathway*. New York, Springer Verlag, 66-116.

Cherry, E. C. (1953). Some experiments on the recognition of speech with one, and with two ears. J. Acoust. Soc. Am. 25, 975-979.

Clarey, J. C., Barone, P., and Imig, T. J. (1992). Physiology of the thalamus and cortex. In: *The mammalian auditory pathway: neurophysiology*. A. N. Popper and R. R. Fay (Eds.), New York, Springer Verlag, 232-334.

Cooke, M., Morris, A., and Green, P. (1996). Recognising occluded speech. Proc. Workshop on the Auditory basis of Speech Perception, Keele, 297-300.

Culling, J. F., and Summerfield, Q. (1995). Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay.. J. Acoust. Soc. Am. 98, 785-797.

Culling, J. F., Marshall, D., and Summerfield, Q. (1998). Dichotic pitches as illusions of binaural unmasking II: the Fourcin pitch and the Dichotic Repetition Pitch. J. Acoust. Soc. Am. 103, 3509-3526.

de Cheveigné, A. (1993). Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. J. Acoust. Soc. Am. 93, 3271-3290.

de Cheveigné, A. (1997). Concurrent vowel identification III: A neural model of harmonic interference cancellation. J. Acoust. Soc. Am. 101, 2857-2865.

de Cheveigné, A. (1998). Cancellation model of pitch perception. J. Acoust. Soc. Am. 103, 1261-1271.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (1997a). Concurrent vowel identification I: Effects of relative level and F0 difference. J. Acoust. Soc. Am. 101, 2839-2847.

de Cheveigné, A., McAdams, S., and Marin, C. (1997b). Concurrent vowel identification II: Effects of phase, harmonicity and task. J. Acoust.Soc. Am. 101, 2848-2856.

de Cheveigné, A., and Kawahara, H. (1998b). Missing feature model of vowel perception. J. Acoust. Soc. Am. 105, 3497-3508.

de Ribaupierre, F. (1997). Acoustical information processing in the auditory thalamus and cerebral cortex. In: *The central auditory system*. G. Ehret and R. Romand (Eds.), New York, Oxford University Press, 317-397.

Durlach, N. I. (1963). Equalization and cancellation theory of binaural masking-level differences. J. Acoust. Soc. Am. 35, 1206-1218.

Durlach, I., and Colburn, H. S. (1978). Binaural phenomena. In: *Handbook of perception*. E.C.Carterette and M. P. Friedman (Eds.), New York, Academic Press, IV, 365-466.

Hartmann, W. M. (1988). Pitch perception and the segregation and integration of auditory entities. In: *Auditory function - neurological bases of hearing*. G. M. Edelman, W. E. Gall and W. M. Cowan (Eds.), New York, Wiley, 623-645.

Heffner, R. S., and Heffner, H. E. (1992). Evolution of sound localization in mammals. In: *The evolutionary biology of hearing*. D. B. Webster, R. R. Fay and A. N. Popper (Eds.), New York, Springer-Verlag, 691-715.

Helfert, R. H., and Aschoff, A. (1997). Superior olivary complex and nuclei of the lateral lemniscus. In: *The central auditory system*. G. Ehret and R. Romand (Eds.), New York, Oxford University Press, 193-258.

Helmholtz, H. v. (1877). *On the sensations of tone.* (English translation A.J. Ellis, 1954). New York, Dover.

Hermansky, H. (1998). "Should recognizers have ears?," Speech Comm. 25, 3-27.

Jeffress, L. A. (1948). A place theory of sound localization. J. Comp. Physiol. Psychol. 41, 35-39.

Irvine, D. R. F. (1992). Physiology of the auditory brainstem. In: *The mammalian auditory pathway: neurophysiology*. A. N. Popper and R. R. Fay (Eds.), New York, Spring Verlag, 153-231.

Jeffress, L. A. (1948). A place theory of sound localization. J. Comp. Physiol. Psychol. 41, 35-39.

Joris, P. X., and Yin, T. C. T. (1998). Envelope coding in the lateral superior olive. III. Comparison with afferent pathways. J. Neurophysiol. 79, 253-269.

Lea, A. (1992), Auditory models of vowel perception. Nottingham unpublished doctoral dissertation.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. Experientia 7, 128-134.

Lippmann, R. P., and Carlson, B. A. (1997). Using missing feature theory to actively select features for robust speechj recognition with interruptions, filtering, and noise., Proc. ESCA Eurospeech, KN-37-40.

Lyon, R. F. (1983-1988). A computational model of binaural localization and separation. reprinted in *Natural computation*. W. Richards (Ed.), Cambridge, Mass, MIT Press, 319-327.

Massaro, D. W. (1990). "Models of integration given multiple sources of information," Psychological review 97, 225-252.

Meddis, R., and Hewitt, M. J. (1992). Modeling the identification of concurrent vowels with different fundamental frequencies. J. Acoust. Soc. Am. 91, 233-245.

Morris, A. C., Cooke, M. P., and Green, P. D. (1998). Some solutions to the missing feature problem in data classification, with application to noise robust ASR., Proc. ICASSP, 737-740.

Oertel, D. (1999). "The role of timing in the brain stem auditory nuclei of vertebrates," Ann. Rev. Physiol. 61, 497-519.

Patterson, R., Anderson, T. R., and Francis, K. (1996). Binaural auditory images and a noise-resistant, binaural auditory spectrogram for speech recognition. Proc. Workshop on the auditory basis of speech perception, Keele, 245-252.

Romand, R., and Avan, P. (1997). Anatomical and functional aspects of the cochlear nucleus. In: *The central auditory system*. G. Ehret and R. Romand (Eds.), New York, Oxford University Press, 97-191.

Rouiller, E. M. (1997). Functional organization of the auditory pathways. In: *The central auditory system*. G. Ehret and R. Romand (Eds.), New York, Oxford University Press, 3-96.

Sabatini, B. L., and Regehr, W. G. (1999). "Timing of synaptic transmission," Ann. Rev. Physiol. 61, 521-542.

Schwartz, I. R. (1992). The superior olivary complex and lateral lemniscal nuclei. In: *The mammalian auditory pathway: neuroanatomy*. D. B. Webster, A. N. Popper and R. R. Fay (Eds.), New York, Springer-Verlag, 117-167.

Summerfield, Q., and Culling, J. F. (1992). Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency. Proc. 124th meeting of the ASA, 2317(A).

Szentágothai, J., and Arbib, M. A. (1975). *Conceptual Models of Neural Organization*. Cambridge. MA., The MIT Press.

Trussel, L. O. (1999). "Synaptic mechanisms for coding timing in auditory neurons," Ann. Rev. Physiol. 61, 477-496.

Warren, R. M. (1996). Processing of speech and some other auditory patterns: some similarities and differences., Proc. Workshop on the auditory basis of speech perception, Keele, 226-231.

Weintraub, M. (1985), A theory and computational model of auditory monaural sound separation. University of Stanford unpublished doctoral dissertation.

Yin, T. C. T., and Chan, J. C. K. (1990). Interaural time sensitivity in medial superior olive of cat. J. Neurophysiol. 64, 465-488.