The auditory system as a separation machine

Alain de Cheveigné
CNRS and ATR-HIP

ABSTRACT

This paper is written from the hypothetical standpoint that the auditory system is designed to *separate* sounds rather than just detect, discriminate, or recognize them. Auditory structures and processing mechanisms are judged on their ability to produce a "separable representation" in which correlates of different sources can be selected or ignored. The cochlear filter is assumed to split acoustic information into band-limited channels, rather than just produce a spectral representation (Fourier transformation). Tonotopy, prevalent throughout the auditory system, is assumed to reflect the need to keep the channels apart, rather than the mere repetition of a spectral representation. Between-channel segregation is supplemented by within-channel segregation based on time-domain processing, both binaural (cross-correlation and equalization-cancellation), and monaural (autocorrelation and harmonic cancellation). Binaural processing accounts for binaural unmasking and certain binaural pitch effects. Monaural processing accounts for F_0-driven segregation, and pitch and timbre perception. An essential ingredient in this hypothesis is "missing-feature theory", that deals with the incomplete patterns produced by the segregation mechanisms. Parts of a pattern are weighted according to their reliability, and missing or unreliable evidence is ignored.

1 The three ages of perception

One can distinguish three levels of perception: *detection*, *discrimination*, and *segregation*, that might correspond to three ages in the evolution of perceptual systems. To illustrate the first (detection), Szentagothai and Arbib (1975) give the example of a primitive fish-like organism with a very simple nervous system. The organism has two fins, one on each side of the body, and two sensors placed on either side of the head. Each sensor is directly connected to the opposite fin by a neuron. When food is sensed on one side, the information is transmitted to the fin on the opposite side. 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*. Szentagothai and Arbib suggest that the same is basically true for higher organisms, with additional levels of inhibition that complexify behavior. In higher organisms, the crossed pathway between brain and body would be a heritage of the crossed nervous system of this primitive organism.

This simple organism can survive in a world where things visible are also eatable. In a world containing predators in addition to prey, a more complex behavior is required. Based on what it sees, the organism must decide whether to activate the contralateral fin to get closer and eat, or else activate 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 could also speak of recognition or identification.

A discriminating organism is more likely to survive than a mere detecting one, but both require that prey and/or predators appear one-by-one. If both predator and prey (or several of each) appear together within the field of vision, the organism won't know how to react. In presence of two prey, the organism may take an intermediate course and miss both (as actually occurs with some birds of prey). To survive in such a densely populated world, the organism must be capable of *segregation*. Segregation is the ability to selectively process perceptual evidence by parts, assigning them appropriately to sources in order to maintain a faithful 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, discrimination or recognition, the entire sound can be attributed to one source. For segregation, the waveform (or perceptual representations derived from it) must be *partitioned* and shared between sources. The partition must precede extraction of source qualities, yet it appears to also depend upon those qualities, a paradox emphasized by Bregman (1990). Possibly because of this conceptual difficulty, classic psychoacoustics has concentrated on detection and discrimination, and only recently has segregation come to the forefront with the ideas of Bregman (1990) and others. Over a century ago Helmholtz (1877) had asked how one hears the quality of an instrument playing among others. However that question was put aside for the following century.

To summarize, segregation is an essential task for survival but harder to account for than "classic" tasks of detection or discrimination. The hypothesis explored here is that the auditory system is in large part designed for that task. An essential element of the argument comes from *missing-feature theory*.

2 Missing-feature theory

Consider a speech recognition system preceded by a "computational auditory scene analysis" (CASA) front-end. Suppose that the input speech is corrupted by noise. The CASA front-end may be successful in suppressing it, but parts of the speech patterns are likely to be suppressed at the same time. If the speech recognition system was trained on complete data, it will behave poorly. Missing-feature theory deals with this situation (Cooke et al., 1996, 1997; Lippmann, 1997; Morris et al., 1998). Several options are available, that can be qualified as "bad", "better", or "optimal". A "bad" option is to set 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 for a system that must resynthesize speech after segregation. However for the purpose of recognition, the "optimal" option is generally to *ignore* the missing data. Interpolation of missing data does not create any new information: it is essentially a principled guess, and as such it may be wrong. Ignoring missing data is a safer course.

Missing-feature theory has been applied to practical applications such as speech recognition and vision (Ahmad and Tresp, 1993) , but its usefulness is wider as an ingredient of perception models. It offers an explanation of the continuity illusion and phonemic "restoration" effects, without the need to postulate perceptual synthesis of low-level correlates. It also offers a framework for cross-modal integration of information, each mode being weighted according to its reliability.

Missing-feature theory is useful in perceptual models of source segregation to handle the incomplete patterns retrieved by a segregation mechanism. Missing or unreliable portions must be labeled as such: that is the responsibility of the segregation mechanism.

3 Is the auditory system designed for segregation?

3.1 Tonotopy

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).

In addition to an orderly distribution of CFs, the widths of tuning curves are often similar to those of auditory nerve fibers, implying rather little convergence between neighboring channels (this is true of some but not all tonotopic representations: others involve both excitatory and inhibitory convergence). ICC for example is divided into laminae, that are stacked in tonotopic order along an axis perpendicular to their plane. *Within* each lamina, neurons differ according to other parameters: fine tuning, bandwidth, best modulation frequency, interaural time difference (ITD) or level difference (ILD) tuning, etc.. Although there is evidence for a regular mapping of some of these parameters, it is clear that they cannot all be distributed independently within the two dimensions of a lamina (Irvine, 1992).

Traditionally, cochlear analysis is assimilated to a Fourier transform, and tonotopically organized neural relays to repetitions of a "spectral representation". The ubiquity of tonotopy is taken as evidence for the importance of spectral coding. Why it must be repeated at every level, however, is not clear. An alternative explanation is that peripheral analysis splits the incoming sound into an array of partly redundant band-limited channels, in order to allow differential weighting of portions of the spectrum according to their reliability, or according to the source that dominates them. This hypothesis, very similar to one proposed by Møller (1977), is explored in this paper.

3.2 Time-domain processing

Auditory-nerve fibers synchronize to the fine structure of stimuli. Measures of synchrony tend to drop above 1-2 kHz, but they remain significant up to 4-6 kHz (9 kHz in the barn owl). The 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 behavioral experiments show that ITDs as small as 6 μs can be exploited (Irvine, 1992). The upper frequency limit of synchrony might reflect a difficulty in coding repeated features at a high rate, in addition to limited temporal resolution per se.

Certain neural hardware seems to be designed for coding temporal information: specialized synapses, large cell bodies, fast membrane potential recovery, etc.. In CN, spherical bushy cells (SBC) are fed by single auditory-nerve fibers via the "endbulbs 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 a small number 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, there are others, such as octopus cells in CN, that 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).

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, in addition to CN and the cochlea (Schwartz, 1992; Romand and Avan, 1997, Helfert and Aschoff, 1997). High-resolution temporal information is thus available at many levels below IC, possibly including levels where synchrony is not measurable: synchrony may be absent in cells that receive synchronized projections, and hard to measure in the projections themselves for technical reasons, but nevertheless it may participate in signal processing at the site where the projections interact.

Whereas MSO is implicated in the time-domain processing of binaural ITDs (Jeffress, 1948; Yin and Chan, 1990), LSO is traditionally assigned the processing of level differences (ILDs), that a-priori do not require fine time resolution. However, if such were the case, the time-specialized circuits that feed LSO from ipsilateral CN and contralateral MNTB would be hard to justify. Recent studies have suggested that LSO plays a role in processing *dynamic* ILDs (onsets), localization via a "negative" version of Jeffress's model, 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 very limited convergence of neighboring CFs in these circuits. SBCs are fed by single AN fibers, GBCs by small groups of presumably similar AN fibers, and MNTB principal cells by single GBC axons. Processing occurs independently for different CFs. It is thus 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., 1998a,b).

The time-specialized CN/MNTB/LSO/MSO circuitry is usually assigned the role of processing binaural ITD and ILD information. This heavy investment is hard to justify for a function that is of secondary importance, and undeveloped in many animals (Heffner and Heffner, 1992). LSO, for example, is little developed in humans. As noted earlier, the circuit has many other projections. It is unlikely that these projections are *all* involved in binaural processing, even if that role is put forward in many studies (no doubt for lack of a better idea). The thesis defended in this paper is that they may be involved in time-domain segregation processes that complement the across-channel segregation supported by tonotopy.

4 Frequency analysis, the "last linear stage"

Various non-linearities are known to exist in the cochlea, but the domain of mechanical vibration is nonetheless likely to have better linearity and dynamic range (in the sense of independent coding of components with different amplitudes) than subsequent neural stages. The cochlea can be seen as a "last linear stage", in which acoustic information is prepared for subsequent analysis in the auditory nervous system (Møller, 1977, 1983).

The role of peripheral analysis in *separating* sounds is evident in masking. Detection and/or discrimination are a direct function of the ability of the system to exclude interference from certain channels responding to the target, and concentrate on those channels and ignore others. The question is usually not addressed explicitly, but channels must somehow be *labeled* for attention or suppression. Channel labeling is explicit in Meddis and Hewitt's (1992) concurrent vowel segregation model, where channels are labeled by dominant periodicity. Earlier attempts along the same line were Lyon's (1983) binaural and Weintraub's (1985) monaural segregation systems. These models are explained in more detail in the next sections.

Meddis and Hewitt's model treats non-selected channels as having a *value* of zero. Missing-feature theory suggests that it is better to give them a *weight* of zero. Because channels are summed before pattern matching, this distinction has no meaning in M&H's model, but it would if the model performed pattern-matching directly on the multichannel ACF array.

Unequal weighting of frequency channels can explain why speech is resistant to severe narrow-band filtering or masking by narrow-band noise (Warren, 1996). In an experiment with concurrent vowel stimuli of same F_0, I found that identification of a vowel was hardly disrupted by the presence of formants of another vowel (de Cheveigné, 1997a). Apparently the presence of the target's formants was treated as evidence for that vowel, while spurious formants of the second vowel were ignored. This can be explained by supposing (a) that spurious formants are labeled as belonging to the second vowel, and (b) that the corresponding channels are ignored in the pattern-matching process that recognizes the first vowel.

5 Binaural analysis

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-labeling, and channel-splitting.

Channel-labeling follows the ideas of Lyon (1983). Lyon used an array of cross-correlation functions, similar to that involved in the Jeffress (1948) localization model. In Jeffress's model, a peak in the cross-correlation array signals the azimuth of a source. In Lyon's system, the peak appears in different positions in different channels, according to which source dominates them. This information is used to *label* channels, and thus separate information belonging to each source. The channel-labeling principle has been used repeatedly in binaural models (for example Patterson et al., 1996). Channel-labeling 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 binaurally (according to a channel-labeling model).

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 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.

The EC and modified EC models involve time-domain interaction of neural signals with high temporal resolution. However their output is usually smoothed and treated as a slowly-varying spectral pattern (residual activity vs channel). This assumption is not necessary: the temporal structure might just as well be conserved at the output, and submitted to additional time-domain processing.

Equalization-and-cancellation is highly effective in theory. It offers infinite noise rejection for a single, well-localized masker, compared with the mere 6 dB boost (for a well-localized target) offered by additive beamforming. The price to pay is *spectral distortion* due to comb-filtering. If τ is the difference in interaural time-of-arrival between target and arrival, and α a factor that represents the combined effects of EC scaling and target ILD, then the target undergoes filtering by a comb filter with the following impulse response:

$$h(t) = \delta(t) - \alpha\delta(t - \tau) \tag{1}$$

The transfer function has zeros at 0 Hz and all multiples of $1/\tau$. Their depth depends on α and is infinite if $\alpha = 1$. Assuming a maximum physiological ITD of 0.7 μs, $1/\tau$ can take any value upwards of about 700 Hz. Such spectral distortion may cause a mismatch in pattern matching, particularly if the zero coincides with an important formant. However, given that the nature of the distortion is known to the system, the mismatch can be eliminated using missing-feature techniques. There is nevertheless a loss of information: spectral patterns differing only at multiples of $1/\tau$ cannot be discriminated.

6 Harmonic analysis

A sound that is periodic (in time), or equivalently harmonic (in frequency) generally 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. In the case of two competing harmonic sounds (two voices), there are potentially two harmonic series to exploit. However it turns out that the intelligibility of a voice does not depend on its own harmonic structure, but only on the harmonic structure of the interference (Summerfield and Culling, 1992; Lea, 1992; de Cheveigné et al., 1995, 1997a,b). In other words, the harmonic structure of interference is exploited to *suppress* it, but the harmonic structure of a target is not exploited to enhance that target. [note: This latter result is certainly counterintuitive, and it contradicts many segregation models. One should not exclude the possibility that the harmonic structure of a target *is* exploited in some way yet to be revealed experimentally.]

Like binaural models, harmonic segregation models can be divided into two classes: channel-labeling and channel-splitting. Channel-labeling based on the ideas of Weintraub (1985), has been recently developed by Meddis and Hewitt (1992). An array of autocorrelation functions (ACF) is calculated, one for each channel. The position of the major peak ("period peak") of the ACF within a channel indicates the period that dominates it, and thus allows the channel to be labeled as belonging to one source or the other. For concurrent vowels, the formant peaks of one vowel often correspond to spectral valleys of the other. When such is the case, channels are easy to assign to vowels, and the model is successful in segregating the vowels.

Channel-splitting is performed in the concurrent vowel identification model of de Cheveigné (1997b). Each channel is processed by a "neural cancellation filter" (de Cheveigné, 1993), tuned to suppress the period of the interference. A model based on this filter accounts for experimental results very well. In particular it explains why F_0-guided segregation is effective even when the amplitude ratio is large (15 to 25 dB), in which case all channels are dominated by one vowel and channel-selection must fail. As was the case for binaural cancellation models, the output of the array of cancellation filters can be processed either as a slowly-varying spectral pattern, or as an array of time-domain patterns.

There is a similarity between models of pitch and models of F_0-guided segregation, whether they are based on autocorrelation (Meddis and Hewitt, 1991 and 1992 respectively), or cancellation (de Cheveigné, 1997b and 1998 respectively). This confirms Hartmann's (1988) suggestion that pitch and segregation are closely related. There are also similarities between models of pitch and localization (Licklider, 1951; Jeffress, 1948), and also between models of monaural (F_0-guided) and binaural segregation. Indeed, Nordmark (1963) has noted strong analogies between pitch and binaural phenomena.

Channel-labeling and channel-splitting models both produce, at their output, patterns that are distorted or incomplete. For the former, any channel attributed to source A is missing for source B. For the latter, spectral distortion affects all channels. Spectral distortion can be described as the effect of filtering with the following impulse response:

$$h(t) = \delta(t) - \delta(t - T_i) \tag{2}$$

where T_i is the period of the interfering voice. The filter has zeros at 0 Hz, and multiples of $1/T$. If the target is also voiced, filter zeros and target harmonics interact to form a sort of "moiré" pattern, equivalent to filtering with the following impulse response:

$$h(t) = \delta(t) - \delta(t - |T_t - T_i|) \tag{3}$$

where T_t is the period of the target voice. These various forms of distortion are known to the system, and can be compensated for by using missing-feature techniques.

7 Lag-domain analysis & vowel timbre

The timbre of a steady-state sound such as a vowel is traditionally attributed to spectral characteristics extracted from a spectral-domain (place) representation. However Meddis and

Hewitt (1992) suggested that vowels could be classified based on the short-lag (τ < 4 ms) portion of the summary autocorrelation function (SACF). This idea was also used in the concurrent vowel-identification model of de Cheveigné (1997b).

A vowel perception mechanism must deal with the problem of *undersampling* of the spectral envelope. Vowel identity depends on the shape of the spectral envelope (that reflects the shape of the vocal tract), and particularly the first two or three formants. However, in the vowel production process this envelope is *sampled*, all the more sparsely as F_0 is high. Undersampling evidently causes a loss of information. It also causes aliasing, that can result in an F_0-dependent distortion of auditory representations, whether they be spectral or temporal.

Aliasing can be avoided in a vowel-perception model based on missing-feature techniques (de Cheveigné and Kawahara, 1998a,b). The model can be formulated in either the frequency or the autocorrelation domain. In the frequency domain, an F_0-dependent weighting function is applied to restrict spectral pattern matching to multiples of F_0. In the autocorrelation domain, an equivalent operation is performed by restricting pattern-matching to portions of the ACF array with lags below $1/2F_0$. Details may be found in de Cheveigné and Kawahara (1998a,b).

This "missing-feature" timbre perception model can be coupled with the previous channel-selection and time-domain segregation models, to form a flexible perception model in which missing-feature techniques may be applied along many dimensions.

Summary and conclusion

In this paper, the structure and behavior of the auditory system were "reinterpreted" as serving the purpose of *segregating* sources. This position may seem rather extreme. It was chosen for rhetoric purposes and should not be taken too dogmatically, but it may nevertheless lead to fruitful insights. For example, it turned out that a mechanism postulated for segregation could also serve for estimation: the cancellation filter used for F_0-guided segregation (de Cheveigné, 1993, 1997) is effective when applied to pitch estimation (de Cheveigné, 1998). It is certainly useful to go beyond the classic view of an auditory system as a mere "estimator" of auditory qualities, or "recognizer" of patterns. It is also good to imagine other roles for cochlear selectivity than mere Fourier Analysis.

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 "The mammalian auditory pathway," Edited by D. B. Webster, A. N. Popper and R. R. Fay, 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," Edited by A. N. Popper and R. R. Fay, 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.
- Cooke, M., Morris, A., and Green, P. (1997). "Missing data techniques for robust speech recognition.", Proc. ICASSP, 863-866.
- 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. (1998a). "Dichotic pitches as illusions of binaural unmasking II: the Fourcin pitch and the Dichotic Repetition Pitch," J. Acoust. Soc. Am. 103,
- Culling, J. F., Summerfield, Q., and Marshall, D. H. (1998b). "Dichotic pitches as illusions of binaural unmasking I: Huggin's pitch and the "Binaural Edge 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. (1997a), "Ten experiments in concurrent vowel segregation," ATR Human Information Processing Research Labs technical report, TR-H-217.
- de Cheveigné, A. (1997b). "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., McAdams, S., Laroche, J., and Rosenberg, M. (1995). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. 97, 3736-3748.

- 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. (1998a), "A model of vowel perception based on missing feature theory," ATR-HIP technical report, TR-H-252.
- de Cheveigné, A., and Kawahara, H. (1998b). "Missing feature model of vowel perception," J. Acoust. Soc. Am. (submitted)
- de Ribaupierre, F. (1997). "Acoustical information processing in the auditory thalamus and cerebral cortex," in "The central auditory system," Edited by G. Ehret and R. Romand, 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," Edited by E.C.Carterette and M. P. Friedman, 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," Edited by G. M. Edelman, W. E. Gall and W. M. Cowan, 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," Edited by D. B. Webster, R. R. Fay and A. N. Popper, 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," Edited by G. Ehret and R. Romand, 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.
- 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," Edited by A. N. Popper and R. R. Fay, 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," Edited by W. Richards, Cambridge, Mass, MIT Press, 319-327.
- Meddis, R., and Hewitt, M. J. (1991). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J. Acoust. Soc. Am. 89, 2866-2882.
- 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.
- Møller, A. R. (1977). "Frequency selectivity of the basilar membrane revealed from discharges in auditory nerve fibers," in "Psychophysics and physiology of hearing," Edited by E. F. Evans and J. P. Wilson, London, Academic Press, 197-207.
- Møller, A. R. (1983). "Auditory physiology," New York, Academic Press.
- Nordmark, J. (1963). "Some analogies between pitch and lateralization phenomena," J. Acoust. Soc. Am.J. Acoust. Soc. Am. 35, 1544-1547.
- 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," Edited by G. Ehret and R. Romand, New York, Oxford University Press, 97-191.
- Rouiller, E. M. (1997). "Functional organization of the auditory pathways," in "The central auditory system," Edited by G. Ehret and R. Romand, New York, Oxford University Press, 3-96.

- Schwartz, I. R. (1992). "The superior olivary complex and lateral lemniscal nuclei," in "The mammalian auditory pathway: neuroanatomy," Edited by D. B. Webster, A. N. Popper and R. R. Fay, 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.
- 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.