Chapter 1

THE CANCELLATION PRINCIPLE IN ACOUSTIC SCENE ANALYSIS

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

Abstract *Cancellation* is a process by which an interfering source (the "jammer") is removed from a mixture of sounds on the basis of its structure. This is part of the task of "scene analysis" that confronts natural organisms and artificial devices. Jammer cancellation is distinct from, and complementary to, target enhancement. Time-domain cancellation filters are distinct from, and complementary to, time-frequency analysis filters. The cancellation principle is probably used by the auditory system to analyze acoustic scenes on the basis of the spatial or harmonic structure of interfering sources. It is related to modern techniques such as ICA (Independent Components Analysis).

Introduction

The acoustic environment is often cluttered. The ears of an organism sample *mixtures* of acoustical waveforms coming from multiple sources, rather than the source waveforms themselves. Making sense of the environment on this basis is a process known as Auditory Scene Analysis, or ASA (Bregman, 1990). If the organism is interested in a particular source (the *target*), the presence of other sources (*jammers*) may interfere with target perception. Unfortunately, perceptual models are generally designed to handle a single isolated source, and extending them to work within a complex environment is a challenge. Similar problems arises when designing an artificial device (such as a speech recognizer) to work in an acoustically cluttered environment.

Cues used by humans have been reviewed by Bregman (1990). Generally speaking, they consist in *regularities* of the sources and/or the scene. These include spatial location (correlation between ears or sensors), periodicity (correlation across time), onset (correlation across frequency channels), familiarity (correlation with predetermined templates or patterns) etc. Artificial systems

have been built that use similar regularities (Cooke and Ellis, 2001). Traditionally, most efforts have concentrated on regularities of the *target* that allow it to be enhanced. This paper describes an approach that instead uses regularities of the *jammers* to suppress them.

Compared to target enhancement, jammer cancellation has two advantages. First, in ideal situations, it provides *infinite* jammer rejection, and thus an infinite SNR improvement. In contrast, even in ideal conditions target enhancement usually offers limited gain (for example 6 dB for a two-microphone delay-and-add beamformer). Second, jammer cancellation works well in situations where SNR is unfavorable, for which segregation is most needed. In those situations, the structure of the jammer is easy to estimate while that of the target is hard to estimate.

Cancellation has at least two weaknesses. The first is that "crosstalk" may arise if the jammer is imperfectly structured and thus incompletely suppressed. Because of crosstalk the target may not be observable within certain temporal intervals or spectral bands. The second weakness is that cancellation may "damage" the target, as a result of suppression of target components that are indistinguishable from the jammers. These two weaknesses are distinct. To be effective, cancellation requires techniques to deal with incomplete target observations and to compensate for target distortion.

Jammer "structure" takes many forms. One or several jammers may be predictable, or periodic, or jammer components may be correlated across several sensors. These basic structures may be extended to include amplitude variation, frequency modulation, moving sources, etc. Each bit of exploitable jammer structure opens a window through which the target can be "glimpsed".

The focus here is mainly on artificial systems (typically automatic speech recognition, ASR), but understanding how the auditory system handles such tasks is also a goal, in itself and as a source of ideas for better algorithms. Conversely, effective algorithms may serve as models to guide our investigation of natural processes.

1. Task and context

The task is to recognize or recover a target source within a noisy environment. For simplicity, suppose that there are only two sources T (the "target") and J (the "jammer") that are observed indirectly via signals X and Y measured from one or two microphones. This structure can be generalized to more sources and/or sensors as needed. Sources and observations are related via a *mixing matrix* that is convolutive: each matrix element is a transfer function (or impulse response) that represent the effects of propagation delay and dispersion from a source to a transducer (Fig. 1.1).

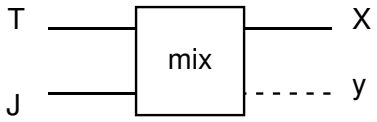

Figure 1.1. Observed signals X and Y are related to target T and jammer J via a mixing matrix.

Two subtasks are of interest. The first is to derive useful information about the *structure* of the scene and/or the sources: intersensor correlation, source fundamental frequencies (F_0s), etc. The second is to recover a "clean" version of the target. Structure estimation is usually a prerequisite of target recovery. It is often possible to derive an approximation T' of the target from the observed signals. T' depends on both target and jammer:

$$T' = f(T) + \epsilon(J) \tag{1.1}$$

Ideally we'd like $f()$ to be identity and $\epsilon()$ to be zero (no distortion and no crosstalk, respectively). Arguably, of these two ideals the latter is the most useful. Whereas target distortion is usually predictable and can be compensated, crosstalk is usually unpredictable and cannot.

Typical application contexts are ASR, conference systems, hearing aids, musical applications (recording, score following, interactive systems), multimedia indexing, etc.

2. Assumptions on source and scene structure

Cancellation is usually applied in the time domain. At each instant t, an estimate of the jammer waveform is *subtracted* from the compound waveform. Three cases are of interest, that differ according to whether the jammer estimate comes from (1) a predetermined template waveform, (2) previous values of the waveform being processed, or (3) the waveform of another sensor.

The first case (subtraction of a waveform template) is ideal but rare. Examples might be a stationary jammer, or the stereotyped waveform of an instrument note, either known beforehand or estimated from the context. It is ideal because subtraction leaves the target undistorted.

The second case is that of a *periodic* jammer $J_t = J_{t-P}$ where P is the period. Suppose that the observed signal is the sum of the target and the jammer $X_t = T_t + J_t$. By subtracting X_{t-P}, the contribution of J is suppressed:

$$T'_t = X_t - X_{t-P} = T_t - T_{t-P}. \tag{1.2}$$

The result T' depends only on T and not on J. It is spectrally distorted as a result of the processing, but jammer rejection is infinite.

The third case is that of multiple sensors in an anechoic environment. Things are a bit simpler if X and Y are rescaled in time and amplitude so that the contribution of J to each is the same: $X_t = J_t + \alpha_x T_{t-\tau_x}$, $Y_t = J_t + \alpha_y T_{t-\tau_y}$. The contribution of J is then suppressed by forming:

$$T'_t = X_t - Y_t = \alpha_x T_{t-\tau_x} - \alpha_y T_{t-\tau_y}. \tag{1.3}$$

The result T' depends only on T and not on J. It is spectrally distorted as a result of the processing, but jammer rejection is again infinite.

These basic cases can be extended. For example the periodic jammer model can be extended to a *variable amplitude* periodic jammer ($J_t = aJ_{t-P}$). For that, Eq. 1.2 is replaced by $T'_t = X_t - aX_{t-P}$. A *variable frequency* jammer can be handled by time warping the observed signal before processing, a moving source by a combination of time warping and gain adjustment, etc. These operations may be performed within bands of a filterbank, with coefficients that vary from band to band.

The basic cases can also be combined (e.g. multiple periodic sources picked up by multiple sensors, etc.). Cancellation fails in two cases: (a) the jammer does not fit any structure model, and (b) it does, but the target fits the same model. The rest of this paper discusses how to handle those cases. Before that, we discuss the issue of *estimation* of the source and scene structure parameters.

3. Estimating source and scene structure

Jammer template. In some cases the jammer waveform can be completely estimated. A simple example is a deterministic stationary jammer such as hum (power frequency harmonics picked up by low-level audio circuits). Granted the mild assumption that the target has intervals of low amplitude, the jammer template can be obtained from a fit to the waveform in those intervals. Granted the further assumption that the jammer is indeed stationary and deterministic, the template is interpolated and subtracted from the entire waveform. The advantage is that the jammer is subtracted, rather than filtered out, and thus there is no spectral distortion of the target. More complex examples are possible but not discussed here.

Periodicity. In other cases the *period* of the jammer, rather than its waveform, can be estimated. Cancellation itself can be used for this purpose. The idea is to search the parameter space of a cancellation filter looking for a *minimum residual output*. For example, to estimate the period of an isolated source the filter defined by Eq. 1.2 is applied and its parameter P is varied until a minimum is found. This principle was applied with success in the YIN method of F_0 estimation (de Cheveigné and Kawahara, 2002). The same principle can be extended to multiple sources (de Cheveigné and Kawahara, 1999; de Cheveigné and Baskind, 2003).

Intersensor delay/attenuation. The jammer waveform may lack structure, but it may contribute to several observations with certain delays and attenuation factors. Again, these factors may be estimated by cancellation. The idea is to search the parameter space of a spatial cancellation filter (null beamformer) looking for a minimum of the residual output. For example, to estimate delay and attenuation of a single source supposing nondispersive propagation, the filter defined by $X_t - \alpha Y_{t-\tau}$ is applied to sensor signals and its parameters α and τ varied until a minimum is found.

The principle can be extended to *dispersive propagation* and *more than two sources/sensors* by splitting the signals over a filterbank and working within narrowband channels. More on this later. From intersensor parameters one can infer source positions (within surfaces of confusion). However spatial parameters are not of direct use for cancellation unless we wish to include spatial constraints, for example within a multimodal system.

Joint estimation. Periods, intersensor parameters, and templates can be estimated *jointly*. In this case, estimation of each aspect of the structure is aided by other aspects. For example F_0 estimation may be aided by spatial structure, and vice-versa. Joint estimation is computationally expensive. This issue is discussed later on.

4. Recovering the target

Supposing the scene fits a structure model, and its parameters are known, a time-domain waveform T' can be obtained according to equations analogous to Eqs. 1.2 and 1.3. This waveform (or its equivalent spectrum) is then fed to a pattern-matching or resynthesis stage, together with structure parameters if needed. As pointed out in Section 2, cancellation allows perfect jammer rejection in ideal conditions. In practice these conditions may arise only within *limited time or frequency intervals*.

5. Local cancellation & missing data

A likely event is that cancellation is possible for a restricted *temporal interval*. For example if the jammer is voiced speech, harmonic cancellation can be applied only during steady-state voiced segments, during which the target may be "glimpsed". Cancellation might also be possible within a restricted *spectral interval*. For example, narrow-band noise may prevent cancellation within some bands. The target is "glimpsed" within the bands that remain. Combining both ideas, one may apply cancellation within a restricted *spectrotemporal region*. Note however that the efficacy of simultaneous bounds in time and frequency is limited by the Gábor relation (Gábor, 1947).

Supposing cancellation is effective only locally, parts of the target will be missing. The parts that remain may nevertheless be sufficient for a task such as pattern-matching (e.g. ASR). *Missing data techniques* have been developed to address this situation (Cooke et al., 1997; Lippmann et Carlson, 1997; Morris et al., 1998). Missing features are either ignored, or (if possible) constrained by bounds derived from the target + jammer mixture. These techniques assume a "mask" to tell them which intervals are missing. In the context of cancellation, the mask is a by-product of the cancellation process.

A second problem is that the target "glimpses" are usually spectrally distorted by the cancellation filters. An option is to compensate by inverse filtering, but a more general solution is to apply similar distortion to the *templates* in the pattern-matching stage. Information needed for that purpose may be available from the cancellation stage. Template (or model) adjustment is not yet common among missing feature techniques (see de Cheveigné, 1993b, for an early attempt).

6. Models

Pattern-matching is a special case of *model fitting*. Once a model is fitted (possibly on the basis of incomplete data) it allows *interpolation*. The models embedded in an ASR system (states, covariance matrices, dictionaries, etc.) can be used in this way. Other useful models are articulatory, multimodal, linguistic, etc. *Redundancy* relations between features may allow accurate interpolation when one feature is missing and the other not.

7. Power and variance partition

Obviously one must know which features are reliable and which are not. This section suggests one possible approach to obtaining this information from the observed signals. The idea is to partition the *power* within a mixture into parts that reflect various sources. This partition is also useful as a partition of the *power spectrum* (thanks to Parseval's relation). A partition of power can also be interpreted as a partition of *variance* (sum of squares). Variance estimates can then be used to parametrize statistical models from which feature reliability can be inferred.

As an example, consider a quasiperiodic jammer J. It is possible to express it as the sum of two signals J' and J'':

$$J'_t = (J_t - J_{t-P})/2, \qquad J''_t = (J_t + J_{t-P})/2 \tag{1.4}$$

If J is purely periodic with period P, then $J' = 0$ and $J'' = J$. J' is nonzero only if J is not perfectly periodic, and in that sense we can call J' the "aperiodic" part of J, and J'' the "periodic" part.

What makes this partition useful is that it is also a partition of power. Defining the local power of a signal X (measured over a window starting at t) as:

$$||X_t||^2 = (1/W) \sum_{j=t+1}^{t+W} X_j^2, \tag{1.5}$$

it is easy to verify that:

$$(||J_t||^2 + ||J_{t-P}|^2|)/2 = ||J'_t||^2 + ||J''_t||^2 \tag{1.6}$$

The term on the left is the average of two estimates of the power of the jammer (over slightly different windows), and the right hand terms are powers of aperiodic and periodic parts respectively. Parseval's relation implies a similar partition of *power spectra*. Spectrally, the partition can be represented by the transfer functions $1 - cos(2\pi fP)/2$ and $1 + cos(2\pi fP)/2$.

In the context of cancellation, J' represents *crosstalk*. If T' is the cancellation-filtered target, the output of the cancellation stage is $T' + J'$. The quality of the recovered target depends on the relative weights of $||T'||$ and $||J'||$. These cannot be observed, but there are several situations where they can be inferred:

1 Jammer properties may be known well enough to put an upper bound on the ratio $||J'||/||J||$. Using the power of the observed signal $||X||$ as a statistically conservative bound on $||J||$, we get an upper bound on crosstalk power $||J'||$. Thanks to Parseval's relation, this reasoning may be applied to each *frequency*.

2 The target too may be periodic. A full analysis is complicated and will be outlined only briefly. Calling P and Q the periods of jammer and target, the observable signal X can be expressed as the sum of four parts:

$$\begin{aligned} X_t^1 &= (X_t - X_{t-P} - X_{t-Q} + X_{t-Q-P})/4 \\ X_t^2 &= (X_t + X_{t-P} - X_{t-Q} - X_{t-Q-P})/4 \\ X_t^3 &= (X_t - X_{t-P} + X_{t-Q} - X_{t-Q-P})/4 \\ X_t^4 &= (X_t + X_{t-P} + X_{t-Q} + X_{t-Q-P})/4 \end{aligned} \tag{1.7}$$

As above, this defines a partition of signal power. The first quantity X^1 is zero iff target and jammer are perfectly periodic (quantities X^2 and X^3 are zero if target or jammer are periodic, respectively). Under certain assumptions X^1 can be used as an estimate of the power that is "unaccounted" for by a sum-of-periodic-signals model, i.e. crosstalk. Again, this reasoning can be applied to each frequency, based on Parseval's relation.

Similar operations can be performed in the multisensor and hybrid cases. Power is defined as a mean sum of squares, and as such it is equivalent to

mean *variance*. Ratios of variance can be interpreted as measuring the uncertainty with which the target is observed withn in each frequency band, at each time frame, and thus the power partition offers the opportunity of interpreting observations according to a statistical model.

8. Relation with auditory models

Barlow (1961, 2001) suggested that the role of sensory relays is to recode incoming patterns in a way that minimizes numbers of neural discharges, and thus metabolic cost, on average. Cancellation fits this description. A "neural cancellation filter" (e.g. de Cheveigné 1993a) minimizes its output for a periodic input, and at the same time characterizes the regularity of the input pattern.

Durlach's (1963) equalization-cancellation (EC) model proposed that patterns from one ear are subtracted from those from the other (after delay and amplitude scaling) to suppress correlates of a spatially localized jammer. Culling and Summerfield (1995, Culling et al., 1998) proposed a "modified EC" model in which such cancellation occurs independently within peripheral filter bands. In this model, EC parameters are determined from information within a band, and may differ from band to band. See also Breebart et al. (2001) and Akeroyd and Summerfield (2000).

A monaural "harmonic cancellation" model was proposed by de Cheveigné (1993a) and found to account for behavioral data on concurrent vowel identification (de Cheveigné, 1997). In particular it accounted for conditions where one vowel is much weaker than the other, for which other explanations fail. A "cancellation model of pitch perception" was proposed by de Cheveigné (1998). A model that explains pitch shifts of inharmonic partials (Hartmann and Doty, 1996) was proposed by de Cheveigné (1999a).

Given the general functional usefulness of cancellation (as argued in this paper) and the fact that some of these models account for effects that no other model accounts for, it is likely that the cancellation strategy is used within the auditory system. Understanding auditory processes is goal that is worthy in itself. It is also a source for insight into effective processing techniques, and a great opportunity for interaction of mutual benefit between scientific and technological fields. To constrain and develop such useful models, there is strong need for more data on natural systems via behavioral, physiological, and imaging techniques.

9. Relation with other processing techniques

Decomposition within the Time-Frequency plane

Many efforts have been devoted to computational models of ASA (e.g. Cooke, 1991; Brown, 1992; Ellis, 1996). A common approach is to assume a *spectrotemporal decomposition* of each sensor signal over a filterbank, *grouping* together of filter bands that that belong to the target, and their *segregation* of these bands from bands that belong to other sources. Bands are assigned according to a time-frequency "map" that looks like a checkerboard.

The idea comes from the ASA rules reviewed by Bregman (1990), themselves based on the principle of peripheral frequency analysis that originated with Helmholz (1877). Strict Helmholtzian doctrine would have had it that the outputs of the bands form collectively a *spectrum* of slowly-varying values (excitation pattern). Recent thinking, both in auditory models and in CASA systems, allows for each band to carry a *temporal* structure, that may be used to decide how the band is assigned. Early examples are the two-channel system of Lyon (1983), that drew on Jeffress's (1948) localization model to segregate bands according to source bearing. Another is the single-channel system of Weintraub (1985) that drew on Licklider's pitch model to segregate bands according to source periodicity. More recent examples are the CASA systems of Cooke (1991), Brown (1992) or Ellis (1996). Decomposition into time-frequency "pixels" is also used in missing-feature techniques (Cooke et al., 1997; Lippmann and Carlson, 1997; Palomaki et al., 2001), statistical methods for time-frequency pixel assignment (Roweis, 2000, 2003), or multiple F_0 estimation (Wu et al., 2003).

There is considerable variety between systems based on time-frequency analysis. Frequency analysis may be performed by a bank of "auditory" filters, by a standard short-term Fourier transform, or by a more exotic time-frequency transform. The output is either a slowly-varying spectrum, or a set of rapidly-varying temporal waveforms filtered from the input waveform. At each instant a band is assigned entirely to a source ("black and white" map) or only partially ("gray-scale" map). Common to all is that band are "atomic" in the sense that they are not analyzed further.

The effectiveness of the time-frequency approach is limited by the Gábor relation: $\Delta f \Delta t \leq constant$. As an example, the response of a 1 ERB wide gammatone filter centered at 1 kHz is still only 20 dB down (1 % power) at 200 Hz away from the peak. Its impulse response is 20 dB down at 6 ms from the time of peak response. Spectral resolution can be improved only at the expense of temporal resolution, and vice-versa, and so jammer rejection cannot be perfect.

Cancellation is complementary with time-frequency analysis. In ideal conditions if offers perfect jammer rejection , but these ideal conditions may pre-

vail only within a limited time-frequency region. Cancellation cannot be subsumed by time-frequency analysis, but the two approaches are complementary and may usefully be combined.

Enhancement

Enhancement is the mirror image of cancellation. Rather than using jammer structure, *target* structure (periodicity, spatial position) is used to enhance a structured target relative to an unstructured background. Enhancement schemes are much more common in the literature than cancellation. However the SNR improvement that they provide is generally limited. For delay-and-add beamforming it is 6 dB for two sensors, and greater improvement requires more sensors. For harmonic enhancement it is 6 dB for a simple comb-filter, and greater improvement requires filters with longer impulse responses (de Cheveigné, 1993a, Appendix A). Cancellation is distinct from (and complementary to) enhancement.

ICA

Independent component analysis and cancellation are related. The objective of ICA is to produce outputs that are statistically independent. This can happen only if each output depends on one source only, a goal that is attained only if contributions of all other sources are suppressed. Thus, the objectives of ICA and cancellation are equivalent, even if the means to attain them are different. The links between ICA and cancellation should be examined more deeply, and it may eventually turn out that ICA and cancellation can be subsumed within a common framework.

It is interesting to note the similarity between Culling and Summerfield's mEC model, and recent frequency-domain ICA techniques (e.g. Anemueller, 2001). Both are congruent with the notion of "local" cancellation described in this paper.

10. Computational considerations

Estimation of structure parameters using cancellation is expensive, because (except in special cases) the parameter space must be searched exhaustively. *Joint estimation* of several parameters is particularly expensive. Techniques to reduce the cost are described in de Cheveigné (2001).

11. Putting it all together

Here are three example scenarii, of varying complexity, of how cancellation might fit together with other techniques to solve a problem of practical interest.

ASR system with single channel input. Cancellation is used for several purposes: (1) for an isolated voice, to provide F_0, F_0-smoothed spectra, and a time-frequency "harmonicity map" as features for ASR, (2) for two concurrent voices, to provide "glimpses" of both voices, together with time-frequency reliability maps for both. These are used by the ASR stages to constrain models of one or more speakers. Spectral distortion caused by cancellation is compensated in the ASR stage by adjusting spectral models.

Active multimodal recording system. A room (conference room or concert hall) is equipped with a distributed network of switchable microphones (or robot controlled microphones) and video cameras. Cancellation is used to analyze the acoustic structure of signals provided by the microphones. The harmonic structure of sources (voices, instruments) is used to facilitate the acoustic analysis. Its result feeds a spatial model that is also informed by video and any other relevant information. The spatial model is used to switch or move microphones, to optimize pickup and segregation of each source of interest, or to produce a visual display of use to the sound engineer. Cancellation analysis reveals that scene structure information is incomplete (for example intersensor correlation may be good only at note onsets, for which the anechoic propagation approximation is good). Incomplete information is interpolated using *missing data techniques* to constrain *models*. Models are then used in the next stage to *interpolate* across missing parts. Models at all stages, including ASR, can be merged and fit jointly (e.g. Nakamura and Herakleous, 2002). On the basis of models, it may be possible to *resynthesize* high quality speech or music sounds (e.g. Kawahara, this volume).

Multimedia indexing and search. A major problem in dealing with massive volumes and fluxes of multimedia data, as they occur today, is indexing and search. The concept of *metadata* has been invented for that purpose. Arguably the most useful kind of metadata are *content-based*, as they are cheap, reliable and ubiquitous (as opposed to text and other manually produced metadata that are expensive and thus often absent). Content-based metadata can be used to map out redundancies (e.g. copies of same data) and constrain other forms of metadata. They are essential for efficient search.

For *mixtures* of audio sources, it would be desirable that the metadata reflect the sources enough to support searching for *individual sources* within the metadata that label the *mixture*. It is usually not possible to split audio data into streams and label each stream. However it is possible to design content descriptors so that they maximize information about component sources. Cancellation is useful for such labeling. As an example, a single channel containing several periodic sources can be processed so as to obtain (a) estimates of the periods, and (b) a periodicity-based decomposition of power and power spectra. It is

not necessary that segregation be perfect: anything that allows pruning of the search space is a sufficiently useful goal.

The power spectrum decomposition is also a decomposition of *variance*, and thus it fits well with statistical models that support hierarchical search (de Cheveigné, 2002). It also fits well with the scalable metadata concept that has been integrated into the audio part of the MPEG-7 standard (de Cheveigné 1999b; ISO/IEC_JTC_1/SC_29, 2001). The additive nature of variance implies that "decomposed" and "standard" descriptions are compatible. Together with the scalability of metadata structures (also based on variance), this ensures interoperability and flexibility of the metadata descriptions.

12. Conclusion

Cancellation is a useful "ingredient" to solve the problems of speech separation and acoustic scene analysis. Other essential ingredients are time-frequency analysis, models, and missing-data techniques. The strength of cancellation is that it can provide, in ideal conditions, infinite jammer rejection. Its weakness is that these ideal conditions may occur only locally, in time and/or frequency, hence the need for models and missing-data techniques. This approach should benefit from future progress in signal processing techniques such as beamforming and ICA, and also from being cast into a systematic probabilistic framework. There are arguments to say that neural processing in natural organisms is in part based on cancellation. More basic knowledge is needed about the nature of these mechanisms, their anatomy and physiology, and the behavior that they allow.

Acknowledgments

Thanks to Pierre Divenyi, Dan Ellis and Deliang Wang for organizing the Montreal workshop, and for providing the stimulation to work on these ideas. Thanks to NSF for funding to support the workshop.

References

Akeroyd, M. A., and Summerfield, A. Q., 2000, A fully-temporal account of the perception of dichotic pitches, Br. J. Audiol. 33(2):106-107.

Anemueller, J., 2001, Across-frequency Processing in Convolutive Blind Source Separation, Oldenberg, unpublished doctoral dissertation.

Barlow, H. B., 1961, Possible principles underlying the transformations of sensory messages. in: Sensory Communication, W. A. Rosenblith, ed., MIT Press, Cambridge, Mass, pp. 217-234.

Barlow, H. B., 2001, Redundancy reduction revisited, Network: Comput. Neural Syst. 12:241–253.

Breebaart, J., van de Par, S., and Kohlrausch, A., 2001, Binaural processing model based on contralateral inhibition. I. Model structure, J. Acoust. Soc. Am. 110:1074-1088.

Bregman, A. S., 1990, Auditory Scene Analysis, MIT Press, Cambridge, Mass.

Brown, G. J., 1992, Computational Auditory Scene Analysis: a Representational Rpproach, Sheffield, Department of Computer Science, unpublished doctoral dissertation.

de Cheveigné, A., 1993a, 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., 1993b, Time-domain comb filtering for speech separation, ATR Human Information Processing Laboratories technical report, TR-H-016.

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., 1999a, Pitch shifts of mistuned partials: a time-domain model, J. Acoust. Soc. Am. 106:887-897.

de Cheveigné, A., 1999b, Scale tree update, ISO/IEC JTC1/SC29/WG11, MPEG99/m5443.

de Cheveigné, A., and Kawahara, H., 1999, Multiple period estimation and pitch perception model, Speech Communication 27:175-185.

de Cheveigné, A., 2001, Correlation Network model of auditory processing, Proc. Workshop on Consistent & Reliable Acoustic Cues for Sound Analysis, Aalborg (Danmark)

de Cheveigné, A., 2002, Scalable metadata for search, sonification and display, Proc. International Conference on Auditory Display (ICAD 2002), 279-284.

de Cheveigné, A., and Kawahara, H., 2002, YIN, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Am. 111:1917-1930.

de Cheveigné, A., and Baskind, A., 2003, F0 estimation of one or several voices, Proc. Eurospeech, 833-836.

Cooke, M. P., 1991, Modeling Auditory Processing and Organisation, Sheffield, Department of Computer Science, unpublished doctoral dissertation.

Cooke, M., and Ellis, D., 2001, The auditory organization of speech and other sources in listeners and computational models, Speech Comm. 35:141-177.

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., Summerfield, Q., and Marshall, D. H., 1998, Dichotic pitches as illusions of binaural unmasking I: Huggin's pitch and the "Binaural Edge Pitch", J. Acoust. Soc. Am. 103:3509-3526.

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

Ellis, D., 1996, Prediction-driven Computational Auditory Scene Analysis, MIT, unpublished doctoral dissertation.

Gábor, D. (1947). "Acoustical quanta and the theory of hearing," Nature 159, 591-594.

Hartmann, W. M., and Doty, S. L., 1996, On the pitches of the components of a complex tone, J. Acoust. Soc. Am. 99:567-578.

ISO/IEC_JTC_1/SC_29, 2001, Information Technology — Multimedia Content Description Interface — Part 4: Audio, ISO/IEC FDIS 15938-4.

von Helmholtz, H., 1877, On the Sensations of Tone (English translation A.J. Ellis, 1885, 1954), Dover, New York.

Hess, W., 1983, Pitch Determination of Speech Signals, Springer-Verlag, Berlin.

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

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

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

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.

Nakamura, S., and Heracleous, P., 2002, 3-D N-Best Search for Simultaneous Recognition of Distant-Talking speech of Multiple Talkers, Proc. IEEE ICMI.

Palomaki, K., Brown, G. J., and Wang, D., 2001, A binaural model for missing data speech recognition in noisy and reverberant conditions, Proc. CRAC (Consistent and Reliable Acoustic Cues) workshop, Aalborg, Danmark.

Roweis, S., 2000, One-microphone source separation, in Advances in NIPS, Edited by M. Press, Cambridge MA, 609–616.

Roweis, S., 2003, Factorial models and refiltering for speech separation and denoising, Proc. Eurospeech.

Weintraub, M., 1985, A Theory and Computational Model of Auditory Monaural Sound Separation, Stanford unpublished doctoral dissertation.

Wu, M., Wang, D., and Brown, G. J., 2002, A Multipitch Tracking Algorithm for Noisy Speech, IEEE Trans. ASSP 11:229-241.