NSF workshop - Perspectives on Speech Separation, Montreal, 31 Oct. - 2 Nov. 2003.

The Cancellation Principle in Acoustic Scene Analysis

Alain de Cheveigné, Ircam-CNRS, 1 place Igor Stravinsky, 75004, Paris, France

Abstract

Cancellation is a process by which an interfering source ("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. 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).

I. Introduction

The acoustic environment is often cluttered. The ears of an organism sample *mixtures* of acoustical waveforms coming from multiple sources. 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 ("target"), the others ("jammers") interfere with target perception. Perceptual models are generally designed to handle a *single* isolated source, and extending them to work within a complex environment is a challenge. The same problems arise 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 either the target or the jammer. These include spatial location (correlation between ears or sensors), periodicity (correlation across time), onset (correlation across frequency channels), etc. Artificial systems have been built that use similar regularities (Cooke and Ellis 2001). Traditionally, models and systems have tended to concentrate on regularities of the *target*. This paper describes an approach that concentrates instead on regularities of *jammers* to suppress them.

Compared to target enhancement, jammer cancellation has two advantages. First, in ideal situations, cancellation provides *infinite* jammer rejection, and thus an infinite SNR improvement. In contrast, 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 that case estimating jammer structure is easy, whereas estimating target structure is hard.

Cancellation has two weaknesses. The first is that the jammer may be imperfectly structured or predictable, and thus incompletely suppressed. This results in "crosstalk". If crosstalk is severe the target may not be observable within certain temporal intervals, and/or spectral bands. The second weakness is that cancellation may "damage" the target. Cancellation requires techniques to deal with incomplete target observations, and to compensate the deleterious effects of target distortion. These two weaknesses are distinct.

Jammer "structure" takes many forms. One or several jammers may be predictable, or periodic, or there may be multiple sensors. These basic structures may be extended to include amplitude variation, frequency modulation, moving sources, etc. Every 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 the task is also a goal, in itself and as a source of ideas for better algorithms. Conversely, algorithms 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 and specifity, suppose two sources T (the "target") and J (the "jammer") that are observed indirectly from signals X and Y provided by one or two microphones. This structure can be generalized as needed to more sources and/or sensors. Sources and observations are related via a *convolutive* mixing matrix. 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).

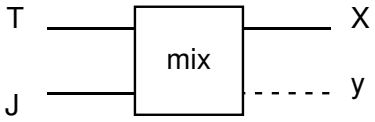

Fig. 1 *Observed signals X (and possibly Y) are related to target T and jammer J via a mixing matrix. The goal is to derive information about the target T.*

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 periodicity (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 usually possible to derive from X or (X,Y) an approximation T' of the target, that depends on both target and jammer:

$$T' = f(T) + \epsilon(J) \tag{1}$$

Ideally we'd like $f()$ to be identity (no distortion) and $\epsilon()$ to be zero (no crosstalk). Of the two ideals, the latter is more useful. Target distortion is typically predictable and can be compensated, whereas crosstalk is usually unpredictable and cannot.

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

2 Assumptions on source and scene structure

Cancellation works 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 template waveform, (2) previous values of the waveform itself, or (3) another sensor.

The first case (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 we observe $X_t = T_t + J_t$ (simple mixture). By subtracting X_{t-P}, the contribution of J is suppressed:

$$T'_t = X_t - X_{t-P}. \tag{2}$$

The result $T'_t = T_t - T_{t-P}$ is a spectrally distorted version of T, but importantly it does not depend on J (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. This contribution is then suppressed by forming:

$$T'_t = X_t - Y_t. \tag{3}$$

Supposing that, after rescaling, D and a are the excess delay and attenuation of T within sensor signal Y with respect to X, the result $T'_t = T_t - aT_{t-D}$ is a distorted version of T. It does not depend on J (jammer rejection is infinite).

These basic cases can be extended. For example a *variable amplitude* periodic jammer ($J_t = aJ_{t-P}$) can be handled by using the formula $T'_t = X_t - aX_{t-P}$. A *variable frequency* jammer can be handled by time warping (the target estimate is then time warped), a moving source with a combination of time warping and gain adjustment, etc. Operations may be performed on bands of a filterbank.

The basic cases can also be combined (e.g. multiple periodic sources picked up by multiple sensors, etc.). The important feature is that jammer rejection is *infinite* each time the jammer fits the structure model. In practice this is likely to not always be the case. 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 *estimation* of the source and scene structure.

3 Estimating source and scene structure

Jammer template. 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, the template is interpolated and subtracted from the entire waveform (for hum this gives excellent results). More complex examples are possible but not discussed here.

Periodicity. Cancellation itself allows estimation. The idea is to design a cancellation filter and search its parameter space for a *minimum residual output*. For example, to estimate the period of an isolated source the filter defined by Eq. 2 is applied and its parameter P is varied until a minimum is found. This principle is 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. Again, cancellation allows estimation. The idea is to design a spatial cancellation filter (null beamformer) and search its parameter space for a minimum residual. 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), but this is not of direct use for cancellation unless we wish to use spatial constraints, for example in 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 uses exhaustive search (with interpolation) of the joint parameter space, and thus is expensive. Computational issues are discussed later on.

4 Recovering the target

Supposing the scene fits the structure model, and parameters are known, a time-domain waveform T' can be obtained according to equations analogous to Eqs. 2, 3. This waveform (together with structure parameters) are then feeded to a pattern-matching or resynthesis stage. Instead of a time-domain waveform, is also possible to derive spectral representations (see below).

As pointed out in Sect. 2, the strength of cancellation is perfect jammer rejection in ideal conditions. Its weakness is that conditions may be less than ideal, or ideal only within certain 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 those segments, the target may be "glimpsed". Cancellation might also be possible within a restricted *spectral interval*. For example, unstructured noise may prevent cancellation within some bands. The target is "glimpsed" within those that remain. One may likewise apply cancellation within a restricted *spectrotemporal region*. However, restrictions in frequency imply smearing in time, and vice-versa, and these limitations must be taken into account (tradeoff between spectrotemporal resolution and maximum allowable jammer rejection).

If cancellation is effective only locally, parts of the target will be *missing*. The parts that remain nevertheless may be sufficient for a task such as pattern-matching (e.g. ASR). *Missing data techniques* have been developped to address this situation (Cooke et al. 1997; Lippmann 1997; Morris et al. 1998). Missing features are either *ignored*, or better (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 is available from the cancellation stage. Template 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. Seemingly trivial *redundancy* relations between features can allow accurate interpolation when one feature is missing and the other not.

7 Power and variance partition

This section suggests how to obtain a reliability mask, and more. 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), and of *variance* (sum of squares).

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

$$J'_t = (J_t - J_{t-P})/2, \qquad J''_t = (J_t + J_{t-P})/2 \tag{4}$$

If J is purely periodic with period P, then $J' = 0$ and $J'' = J$. J' is non-zero 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 power of a signal X (measured over a window starting at t) as:

$$||X_t||^2 = (1/W) \sum_{j=t+1}^{t+W} X_j^2, \tag{5}$$

it is easy to verify that:

$$(||J_t||^2 + ||J_{t-P}|^2|)/2 = ||J'_t||^2 + ||J''_t||^2 \tag{6}$$

The term on the left is an estimate of the power of the jammer (calculated over two windows and then averaged), and the right hand terms are powers of aperiodic and periodic parts respectively. Parseval's relation implies a similar partition of *power spectra*. This is extremely useful. In the spectral domain, the power spectrum is weighted by two complementary functions: $1 - cos(2\pi fP)/2$ and $1 + cos(2\pi fP)/2$ respectively.

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 $||J'||$. Obviously 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:

$$\begin{aligned}
X_t^1 &= (X_t - X_{t-P} - X_{t-Q} + X_{t-Q-P})/4 \\
X_t^2 &= (X_t + X_{t-P} - X_{t-Q} - X_{t-Q-P})/4 \\
X_t^3 &= (X_t - X_{t-P} + X_{t-Q} - X_{t-Q-P})/4 \\
X_t^4 &= (X_t + X_{t-P} + X_{t-Q} + X_{t-Q-P})/4
\end{aligned} \tag{7}$$

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 it can be used to infer the power that is "unaccounted" for, i.e. crosstalk. Again, this reasoning can be applied to each frequency.

Power is defined here as a mean sum of squares. As such it is equivalent to mean *variance*. The partition can be interpreted as measuring the uncertainty with which the target is observed (in each frequency band, at each time frame). This may offer the opportunity of interpreting observed data according to a statistical model. Similar operations can be performed in the multisensor and hybrid cases.

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 the number of neural discharges (and thus metabolic cost) on average. Cancellation fits this description, as a "neural cancellation filter" minimizes its output, 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 (EC parameters differ from band to band, and are determined from information within a band). See also Breebart et al. (2001) and Akeroyd and Summerfield (2000).

A monaural "harmonic cancellation" model was proposed by de Cheveigné (1993a). It was found to account for behavioral data on concurrent vowel identification, in particular conditions where one vowel is much weaker than the other for which other explanations fail (de Cheveigné 1997). 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 functional power 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 a worthy goal of itself, as a source for insight into effective processing techniques, and as a great opportunity for interaction of mutual benefit between scientific and technological fields. There is great need for more data on natural systems via behavioral, physiological, and imaging techniques.

9 Relation with other techniques

A Decomposition within the Time-Frequency plane

A common approach is to assume *decomposition* of each sensor signal over a filterbank, *grouping* together of filter bands that that belong to the target, and their *segregation* from channels that belong to other sources. Channels are assigned according to a time-frequency "map" that looks like a checkerboard.

This approach is common in Computational Auditory Scene Analysis (CASA) systems. 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 pattern across channels forms a *spectrum* of slowly-varying values (excitation pattern). Recent thinking, both in auditory models and in CASA systems, allows for each channel to carry a *temporal* structure, that may be used to decide how the channel is assigned. Early examples are the two-channel system of Lyon (1983), that drew on Jeffress's localization model to segregate channels according to source bearing. Another is the single-channel system of Weintraub (1985) that drew on Licklider's pitch model to segregate channels 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 a slowly-varying spectrum, or a set of rapidly-varying temporal wavefoms filtered from the input waveform. At each instant a channel is assigned entirely to a source ("black and white" map) or only partially ("gray-scale" map). Common to all is that channels 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. The 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 prevail only within a limited time-frequency region (or parameters might vary from region to region). Cancellation cannot be subsumed by time-frequency analysis.

B Enhancement

Rather than jammer structure, it us commonly proposed to use *target* structure (periodicity, spatial position) to enhance a target relative to an unstructured background. The SNR improvement 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.

C 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. That goal is attained only if contributions of all other sources are suppressed. Thus, the result is the same as that aimed at by cancellation, but the means are different. The links between ICA and cancellation should be examined more deeply. It may eventually turn out that ICA can subsume cancellation (i.e. find any solution that cancellation can find).

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 a three example scenarii, some simpler than others, of how cancellation might fit together with other techniques to solve a problem.

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*. For example, a simple model of a source might say that it does not move. Models may also be used in the next stage to *interpolate* across missing parts, in the event that the system was incapable of recovering them. 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 workshop).

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*: they are cheap, reliable and ubiquitous (as compared to text and other manually created metadata), and solve problems such as mapping out redundancies (e.g. copies of same data) that 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 not possible to split data into streams and label each stream. However it is possible to design content descriptors so as to maximize information about component sources. Cancellation allows precisely such labeling. As an example, a single channel containing several periodic sources can be processed so as to obtain (a) estimates of the periods, (b) a periodicity-based decomposition of power and power spectra. It is not necessary that segregation be perfect: enough to allow 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 both interoperability, and the potential to reduce storage cost of descriptions as needed.

12 Conclusion

Cancellation is an essential "ingredient" to solve the problem of speech separation. 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 ideal conditions may occur only locally, in time and/or frequency. Hence the need for models and missing-data techniques. This approach should benefit greatly from 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 cancellation. More 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 this 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.

Anemüller, J. (2001), "Across-frequency processing in convolutive blind source separation," Oldenberg unpublished doctoral dissertation.

Barlow HB (1961) Possible principles underlying the transformations of sensory messages. In Rosenblith WA (ed Sensory Communication. Cambridge Mass: MIT Press, 217-234.

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

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

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.

Brown, G. J. (1992), "Computational auditory scene analysis: a representational approach," 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 technical report.

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), Kyoto (June 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.

Cooke, M. P. (1991), "Modeling auditory processing and organisation," Sheffield, Department of Computer Science unpublished doctoral dissertation.

Cooke, M., Morris, A., and Green, P. (1997). "Missing data techniques for robust speech recognition," Proc. ICASSP, 863-866.

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

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.

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

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.

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

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. (1984). "Computational models of neural auditory processing," Proc. IEEE ICASSP, 36.1.(1-4).

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. (2003). "Factorial models and refiltering for speech separation and denoising," Proc. Eurospeech.

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

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. "A Multipitch Tracking Algorithm for Noisy Speech," IEEE Trans. ASSP 11, 229-241.

Hess, W. (1983). "Pitch determination of speech signals," Berlin, Springer-Verlag.