version 20220213

In-Channel Cancellation

Alain de Cheveigné, CNRS/ENS

Abstract

A model of early auditory processing is proposed in which each peripheral channel is processed by a delay-and-subtract cancellation filter, tuned independently for each channel with a criterion of minimum power. For a channel dominated by a pure tone or a resolved partial of a complex tone, the optimal delay is its period. For a channel responding to harmonically-related partials, the optimal delay is their common fundamental period. Each peripheral channel is thus split into two subchannels, one cancellation-filtered and the other not. Attention can be directed to either depending on the task. The model is applied to explain the masking asymmetry between pure tones and noise: a noise target masked by a tone is more easily detectable than a tone target masked by noise. The model is one of a wider class of models, monaural or binaural, that cancel irrelevant stimulus dimensions so as to attain invariance to competing sources. Similar to occlusion in the visual domain, cancellation yields sensory evidence that is incomplete, thus requiring Bayesian inference of an internal model of the world along the lines of Helmholtz's doctrine of unconscious inference.

Introduction

A pure tone masks a narrowband noise probe less effectively than a narrowband noise masks a tone probe (Helman 1972, Hall 1997). The difference in masking can reach 20 dB or more. This implies a stark deviation from the standard “power spectrum model” of masking according to which a sound is detectable if it its power exceeds that of the masker within at least one peripheral frequency channel (Moore 1995). Similar masking differences have been observed between harmonic and inharmonic (or noise-like) maskers, as reviewed by de Cheveigné (2021): a harmonic masker is less effective than an inharmonic or noise-like masker

of similar spectral envelope, suggesting an unmasking mechanism applicable to harmonic maskers.

Analogous masking differences have been observed in the spatial domain: an inter-aurally correlated masker (as produced by a spatially localized interfering source in an anechoic environment) is less effective than a masker that is inter-aurally uncorrelated. Binaural unmasking is well accounted for by Durlach's (1963) equalization-cancellation (EC) model that assumes that neural signals from both ears are differentially scaled and delayed (equalization), and then subtracted centrally (cancellation). Durlach's (1963) model assumed, tacitly, that the same parameters (delay and scale factor) are applied to every peripheral channel. More recently, “modified EC” models have been proposed that allow for different equalization parameters in each channel (Culling and Summerfield 1994; Breebart and Kohlrausch 2001, Ackeroyd 2004). These modifications are needed to explain unmasking with maskers crafted to have different binaural properties (e.g. different interaural delay) in different frequency bands.

Analogous to Durlach's EC model, the “harmonic cancellation” model of de Cheveigné (1993, 2021) assumes that a neural signal is delayed and subtracted from the non-delayed signal, with a delay equal to the period of an interfering sound. Correlates of a periodic (harmonic) masker are thus suppressed, explaining its reduced masking power. While the original model assumed the same delay in every peripheral channel, that assumption is relaxed here, taking inspiration from the modified EC model. A different delay in each channel allows the model to explain unmasking observed with inharmonic stimuli obtained by stretching, shifting, or jittering the partials of a harmonic complex (as reviewed in de Cheveigné 2021).

However, the in-channel cancellation model goes beyond merely extending the harmonic cancellation model to allow for a different “local fundamental period” in each frequency band. It is also applicable to maskers that are spectrally sparse, so that some individual channels are dominated by a single sinusoidal component of the masker. The simplest example a pure tone, a focus of this paper.

The hypothesis of an automatically-tuned filter in each peripheral channel opens some interesting perspectives. First, it highlights the role of time-domain signal processing in the auditory brainstem, and implies that auditory frequency selectivity is not entirely determined by cochlear frequency selectivity. The idea of a “second filter” dates back to Huggins and

Licklider (1951), more recent incarnations being the lateral inhibitory network (LIN) of Shamma (1985) or the phase-opponency model of Carney et al (2002).

Second, it emphasizes the role of *invariance* as a goal of auditory processing, and cancellation as a fundamental operation to achieve that goal. While cancellation might ensure invariance with respect to the background, the integrity of the target is not guaranteed: certain features may be absent or distorted. Thus, cancellation needs to work hand-in-hand with a Helmholtzian process that can fit an incomplete or distorted target representation to an internal model (de Cheveigné 2021).

Third, a by-product of the fitting process is an estimate of the *period* that dominates each channel. Whereas the original harmonic cancellation model offered a single period estimate, a potential cue to pitch (see de Cheveigné 1998), the in-channel version offers multiple local estimates. A stimulus for which period estimates differ across channels (e.g. an inharmonic complex) typically lacks a clear overall pitch. However, listeners may still be sensitive to *pitch change* (Demany and Ramos 2005, Popham et al 2018) that this model might help explain.

An earlier explanation of the tone/noise or harmonic/inharmonic masking difference is that it is easier to detect a probe on the background of a smooth (unmodulated) pure-tone or harmonic complex than on a rough (modulated) noisy or inharmonic background. The in-channel model can be seen either as an *alternative* to that explanation, or as a sensitive mechanism to detect departure from smoothness as required to *implement* that explanation.

Methods

The model

A simplified linear model of cochlear filtering is used to derive qualitative predictions. For simplicity and clarity, effects of non-linear transduction, compression, and stochastic neural coding are not considered. The initial stage of auditory processing is modeled as a linear filter bank, followed by a delay-and-subtract cancellation filter, with a delay estimated automatically and independently in each channel. The auditory brain is assumed to have access to the original signal in each peripheral channel, the cancellation filter output, and the delay estimate. Details of the simulation are as follows.

Filters

Cochlear filtering is modeled using a gammatone filter bank (Holdsworth et al. 1988, Slaney 1993) with characteristic frequencies distributed uniformly on a scale of equivalent rectangular bandwidth (ERB), with the bandwidth of each channel set to one ERB according to estimates of Moore et al. (1983). Transfer functions of selected channels are plotted in Fig. 1, scaled so that their peak gain is one (0 dB). Plotted on a linear frequency scale, filters appear wider at high than low CF: bandwidth is roughly proportional to CF above 1 kHz, and roughly uniform over the lowest CFs. Each channel attenuates all but a narrow frequency region, so its output is relatively insensitive to the presence of signal features outside that region. However, attenuation is not infinite at any frequency.

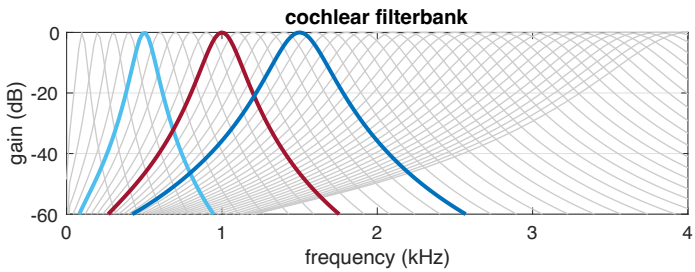

Fig. 1. Transfer functions of selected channels of the gammatone filter bank. Channels with CF=0.5, 1 and 2 kHz are highlighted.

This selectivity can serve to “protect” a narrowband target sound from a competing masker. This is illustrated in Fig. 2 (top) for a 1 kHz probe mixed with a wideband masker of equal RMS. Channels near 1 kHz are purely dominated by the probe, so the presence or absence of the probe can be detected despite the masker. In another example (Fig. 2, middle), individual low-rank partials of a 200 Hz harmonic complex tone are isolated by channels with CFs near each partial's frequency. Two ranges of CFs can be distinguished. Below ~1 kHz (5th harmonic), each partial appears to be perfectly isolated within a subset of channels close to its frequency (intermediate channels respond to at most two neighboring partials). Above ~1 kHz, individual partials are less perfectly isolated (the peaks of the red curve do not reach one). All channels tend to respond to more than two partials (the dotted line represents the power of the third-strongest partial).

This parallels the classic distinction between “resolvable” and “unresolvable” partials of a complex (Moore and Gockel 2011), which is often attributed to the presence or absence of

salient ripples in the excitation pattern (e.g. Fig. 1 in Moore and Gockel 2001). Figure 2 (middle) of the present paper suggests an alternative account: a partial is resolvable if it dominates at least one peripheral channel (i.e. the power of all other components is below some threshold). This might, for example, be a condition for the estimation of that partial's frequency based on time-domain neural patterns (Srulovicz and Goldstein 1983).

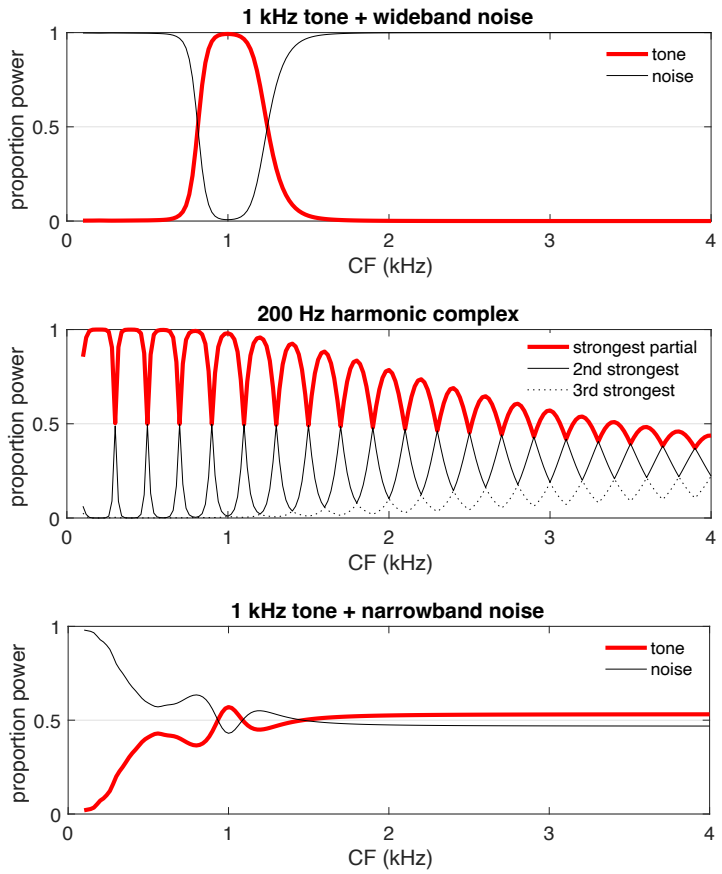

Fig. 2. Proportion of power at the output of a cochlear filter bank as a function of CF for different parts of a stimulus. Top: mixture of a 1 kHz pure tone and wideband noise with equal RMS amplitude (over a 20 kHz bandwidth), red: proportion of power attributable to the pure tone, black: proportion attributable to the noise. Middle: 200 Hz harmonic complex tone. The red line represents the proportion of power attributable to the partial that dominates each channel. Continuous black: proportion attributable to the second strongest partial, dotted black: proportion attributable to the third-strongest partial. Bottom: mixture of a 1 kHz pure tone and a 0.5 ERB wide band of noise centered on 1 kHz, with equal RMS bandwidth.

These first two examples illustrate situations where peripheral filtering can help distinguish a probe sound from a competing background. In contrast, Fig. 2 (bottom) shows a situation where it is of little avail. The stimulus here is a mixture of a pure tone at 1 kHz and a narrowband noise centered at 1 kHz with 0.5 ERB bandwidth (~60 Hz) and equal RMS amplitude. No channel is clearly dominated by the pure tone, which leads us to expect high thresholds for detecting a tone probe in a narrowband noise (or vice-versa). Here, peripheral filtering offers little benefit.

The second ingredient of the model is a cancellation filter, modeled as a simple delay-and-subtract filter with impulse response $h(t) = \delta(t) - \delta(t+T)$ (Fig. 3 left). Its transfer function has zeros at all multiples of $f = 1/T$ (Fig. 3 right), implying that attenuation is *infinite* at those frequencies.

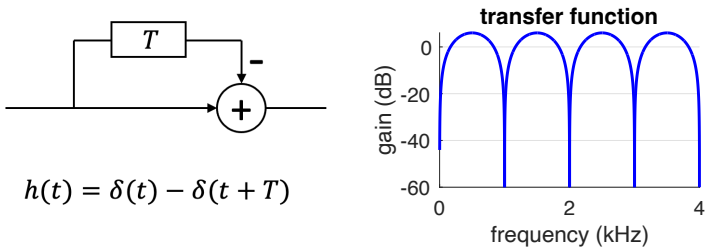

Fig.3. Left: schema and impulse response of a cancellation filter. Right: transfer function for T=1 ms.

If the cancellation filter is applied to the output of a gammatone filter channel, the transfer function of the cascade is the product of their transfer functions, as illustrated in Fig. 4 for three channels with CFs circa 1 kHz. The compound filter inherits spectral filtering properties of its two parts: in addition to reduced gain at frequencies remote from its CF, it has *infinite* attenuation at 1 kHz. The output of this compound filter is invariant to the presence, or absence, of spectral power at precisely 1 kHz.

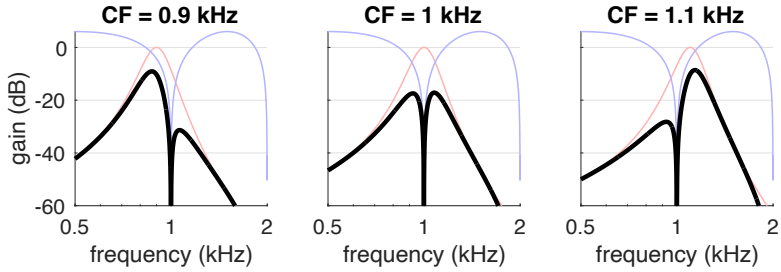

Fig. 4. Transfer function of the cascade of a gammatone filter and a cancellation filter with a delay parameter T=1 ms for three CFs (as indicated). The thin colored lines are the transfer functions of the component filters (red: gammatone, blue: cancellation).

Finding T

The in-channel cancellation filter requires an estimate of the delay T_k in each channel k. Here, we make the important assumption that T_k is determined automatically within each channel, on the basis of the signal within that channel. This is achieved as schematized in Fig. 5, by searching for the non-zero value of τ that minimizes the power at the output of the cancellation filter relative to its input. This is analogous to well-known techniques for period estimation such as AMDF (average magnitude difference function, Ross 1976) or YIN (de Cheveigné and Kawahara 2002). For a periodic sound, the minimum ratio is zero. In the presence of a probe it is non-zero, but the estimate remains accurate as long as the probe is weak (which should be the case at detection threshold).

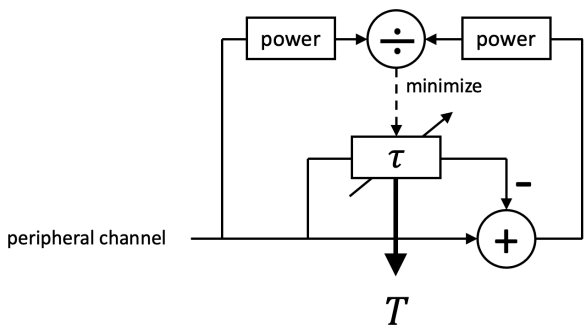

Fig. 5. Automatic determination of the delay T_k within each channel k. The delay parameter τ is varied in search for the smallest ratio of cancellation filter output power to input power.

A couple of details need elaborating. First, input and output power must be smoothed to get a stable estimate of their ratio. The smoothing window must be long enough to suppress

irrelevant fluctuations, yet short enough to track period changes in a non-stationary signal. In the simulations the window size was set to 30 ms. Second, for a purely periodic signal (a pure tone or harmonic complex), the output/input ratio is zero for the period and *all its multiples*, implying an infinite number of equally valid candidates for T_k. For definiteness, it may be convenient to choose among candidates the first for which the power ratio is below some threshold θ (see de Cheveigné and Kawahara 2002 for the rationale and more details). This parameter may affect T_k but it has little impact on the ability of the filter to suppress the background, as reported below.

Figure 6 plots the inverse of the automatically-determined delay T_k within each channel as a function of CF for three stimuli: a 1 kHz pure tone (left), wideband noise (center), and narrowband noise (right). The parameter θ was set to 0 in each case. For the pure tone, the estimate is $1/T_k$ = 1 kHz within all channels. For white noise, it is close to CF for every channel, suggesting that it reflects the noise-excited ringing pattern within each channel. For the narrowband noise, the inverse estimate follows CF at low CFs and stabilizes near 1 kHz at high CFs. These examples illustrate how the delay T_k is automatically chosen for three common stimuli. Additional examples are provided later on.

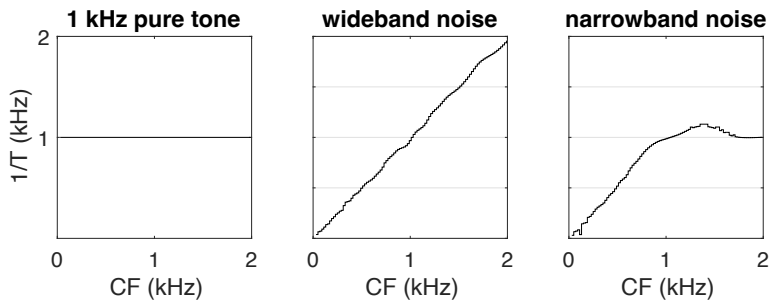

Fig. 6. Inverse of estimated delay T for each channel as a function of its CF for three stimuli. For the pure tone (left), all channels yield $1/T = 1$ kHz. For white noise (center), the value is approximately $1/T$ =CF. For narrowband noise (right) it follows $1/T$ =CF at low frequencies and is close to $1/T = 1$ kHz at high CFs.

For each peripheral channel, the auditory system is assumed to have access to both the cancellation-filtered and unfiltered output, and can attend to either depending on the task. It also has access to the automatically-estimated delay parameter T_k for each channel.

Results

Response to pure tone and narrowband noise

For a pure tone, assuming the delay T_k is accurately chosen, the output of the cancellation filter should be zero. Actually, this is not quite correct for a stimulus of finite length: there is a "glitch" at both onset and offset, but the part intermediate between glitches is zero (Fig. 7 top right). The same is not true for narrowband noise (Fig. 7 bottom right). For that stimulus, the compound filter fails to cancel any portion. The narrowband noise was produced by filtering wideband noise with a gammatone filter centered at 1 kHz with bandwidth 0.5 ERB (roughly 60 Hz), and then shaping it temporally with a 500 ms window with 10 ms raised-cosine onset and offset ramps. In both examples, the delay T_k was chosen automatically as described earlier.

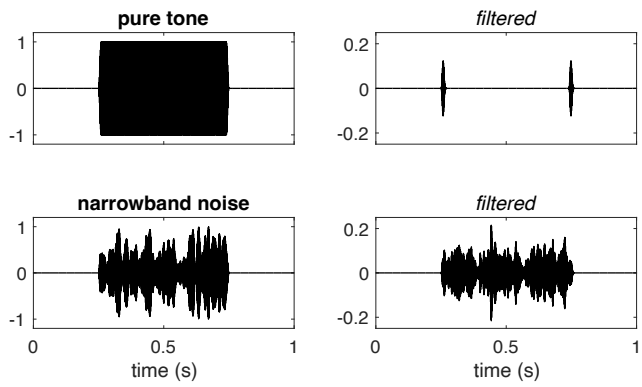

Fig. 7. Effect of the compound filter within the CF=1 kHz channel for two stimuli. Top: 1 kHz pure tone, bottom: narrowband noise centered on 1 kHz (width 0.5 ERB ≈ 60 Hz). Left: stimulus, right: compound filter output. Stimuli are 500 ms in duration and shaped with 10 ms raised-cosine onset and offset ramps.

Figure 8 shows the spectro-temporal excitation pattern across peripheral filter channels for the same two stimuli before (left) and after (right) the cancellation filter. The spectro-temporal pattern was calculated by averaging instantaneous power over a 30 ms window within each channel of the gammatone filter bank For the pure tone, cancellation suppresses

the response of all channels for most of the duration of the stimulus (top right), but the same is not true for narrowband noise (bottom right).

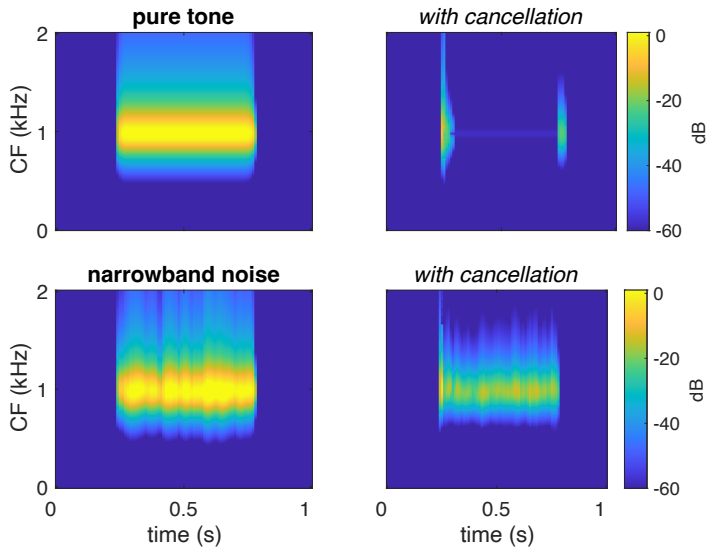

Fig. 8. Spectro-temporal excitation patterns before (left) and after (right) cancellation. For the pure tone (top), the interval between onset and offset is almost perfectly quiescent after cancellation (top right), in contrast to the narrowband noise (bottom) for which excitation persists throughout (bottom right). The auditory brain is assumed to have access to excitation patterns with and without cancellation.

Pure tone/noise masking difference

Figure 9 is similar to Fig. 8, with the addition of a 30 ms probe temporally centered on the tone (top) or narrowband noise (bottom) considered as maskers. The amplitude of the probe was 10 dB below that of the maskers. Its presence is not visually detectable in the absence of cancellation (left). With cancellation, it is prominent for the pure tone masker (top right) but not the narrowband noise masker (bottom right).

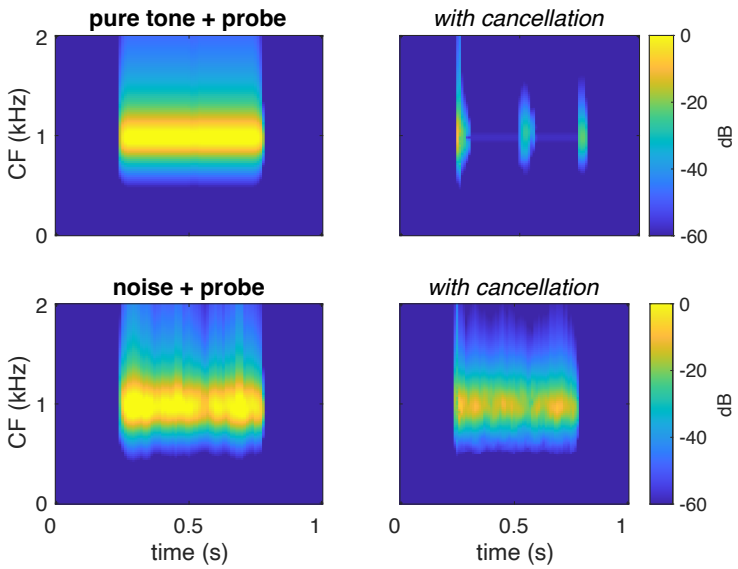

Fig. 9. Spectro-temporal excitation patterns before (left) and after (right) cancellation for a probe added to a masker. The masker was a 1 kHz tone (top) or a narrowband noise of width 0.5 ERB (bottom). The probe was a 30 ms pulse of narrowband noise temporally centered within the masker, with an amplitude -10 dB relative to the masker.

This is qualitatively consistent with the observation, made by Helman (1972) and others, that a pure tone masker is much less potent than a narrowband masker of similar amplitude. The demonstration that this effect is an emergent property of in-channel cancellation is the foremost result of this paper. The rest of this section examines a wider range of examples.

Smoorenburg's sweep tone masker

Smoorenburg and Coninx (1980) found that a pure-tone masker was up to 20 dB less potent than a sweep tone masker when the probe was a short pure tone with a frequency that matched the frequency of the pure tone masker (or the instantaneous frequency of the sweep masker at the probe's position). If anything, the standard power-spectrum model of masking would predict *less* masking for the sweep because it contributes only briefly to power within the channel occupied by the probe.

Figure 10 shows the excitation pattern across channels after cancellation filtering for a 1 kHz pure tone probe temporally centered on a 1 kHz pure tone masker (left) or a sweep-tone masker (right). The sweep tone frequency was swept logarithmically from 500 Hz to 2 kHz in 0.5 s (sweep rate 4 octaves/s). In contrast to the case of a pure tone masker (Fig. 9 left), for the sweep tone the presence of the probe is barely perceptible visually after cancellation

(Fig. 9 right). This is qualitatively in agreement with the stronger masking observed by Smoorenburg and Coninx (1980) for a sweep relative to a pure tone masker

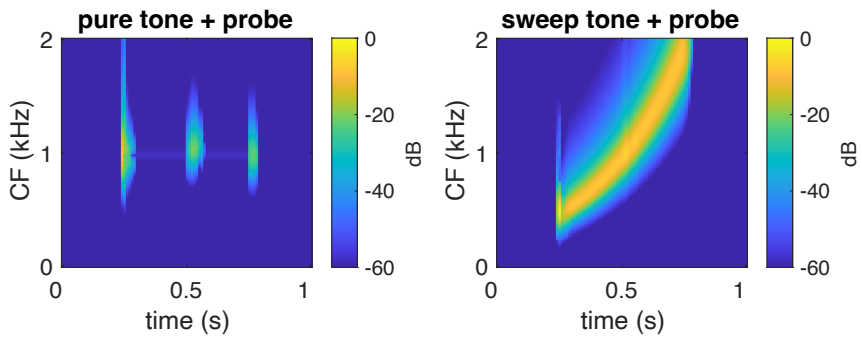

Fig. 10. Spectro-temporal excitation patterns before (left) and after (right) cancellation for a short pure tone probe added to a sweep-tone masker. Cancellation allows the probe to emerge (visually) from the pure tone but not the sweep.

Harmonic and inharmonic maskers

Figure 11 shows the inverse delay parameter $1/T_k$ estimated within each channel as a function of its CF for a 200Hz harmonic complex (left) or an inharmonic complex obtained by jittering each partial of a 200 Hz harmonic complex by a random amount drawn from $[-100, 100]$ Hz (right). For the harmonic complex, at CFs below about 1 kHz the inverse delay matches either the frequency of an individual partial, or the fundamental frequency (200 Hz). For higher CFs, $1/T$ follows CF approximately. For the inharmonic complex (right), for CFs below about 1 kHz, $1/T_k$ appears to follow either the frequency of an individual partial, or a lower frequency that likely corresponds to a common subharmonic of neighbouring partials. For higher CFs, $1/T_k$ tends to follow CF (with some glitches). These estimates were obtained with a threshold parameter $\theta = 0.1$ (see Methods). For a smaller value of θ, the delay estimate is less likely to follow the period of individual partials, and more likely to follow the fundamental period of the harmonic tone, or a "local fundamental" of the inharmonic tone.

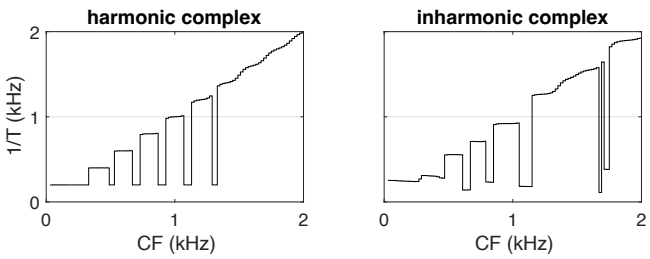

Fig. 11. Inverse of the delay parameter estimated within each channel as a function of its CF, for a harmonic complex (left) or an inharmonic complex (right).

Figure 12 shows spectro-temporal excitation patterns before (left) and after (right) cancellation for the 200 Hz harmonic complex tone (top), the inharmonic complex (middle) and white noise (bottom). For the harmonic complex, cancellation effectively suppresses the excitation pattern over most of the duration of the stimulus (top right). For the inharmonic complex, cancellation is less effective (middle right) than for the harmonic complex, but more than for a noise-like stimulus (bottom right). This may explain why inharmonic maskers are less deleterious than noise-like maskers of similar spectral shape, albeit more than harmonic maskers.

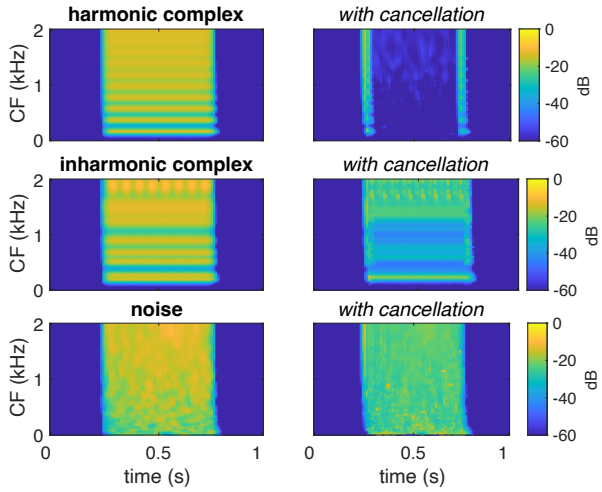

Fig. 12. Spectro-temporal excitation patterns before (left) and after (right) cancellation for three different stimuli as labeled (see text). Cancellation is most effective for the harmonic complex (top) but there is some suppression also for the inharmonic complex (middle).

Discussion

The in-channel cancellation model offers a simple explanation for the asymmetry of masking between pure tone and noise, and extends a previous model (harmonic cancellation) as an explanation of the relatively weak masking power of harmonic and spectrally sparse maskers relative to noise-like maskers of similar spectral content. The results presented here,

based on a simplified linear approximation of cochlear and neural signal processing, are qualitative. Whether they they translate into a quantitative fit with behavioural results awaits more detailed modelling, food for future studies. This section discusses some limits of the model, and speculates on its possible significance as one of a wider class of sound processing mechanisms available to the auditory brain.

The tone/noise asymmetry

The observation that a pure tone is a less potent masker than a narrow-band noise of same intensity (Helman 1972) clashes with the widely accepted power spectrum model of masking according to which a probe is heard as soon as its power reaches some proportion k of the power of the masker within at least one peripheral channel (Moore 1995). As argued by Hall (1997), the mismatch with the power model suggests an additional unmasking mechanism, for which the in-channel cancellation model offers a putative account.

The tone/noise masking asymmetry has also been explained by arguing that the addition of a probe to a pure tone introduces modulation cues that are easier to detect on a smooth background (pure tone) than on an already-modulated background (narrowband noise) (Verhey 2002). The in-channel cancellation model can be seen as an alternative to that explanation, or as an *implementation* of it: in-channel cancellation is arguably a sensitive and expedient way to detect period-to-period fluctuations within each channel. Hall (1997) and Verhey (2002) found reduced masking of a narrowband noise probe by a narrowband noise masker with smaller bandwidth. Verhey (2002) explained these results on the basis of a modulation filterbank (Dau et al 1991): the addition of a probe with a wider bandwidth excites modulation filters of higher modulation frequency than those excited by the masker alone. Using the in-channel cancellation model, we could likewise invoke the fact that a narrowband noise is more *approximately periodic* that a wide-band noise. Moore et al (1998) argue that additional detection cues may need to be taken into consideration, such as distortion products, beats, or dip listening.

Is the model applicable to speech?

Figure 13 (left) shows the spectro-temporal excitation pattern in response to a short speech phrase ("Wow Cool!"). Fig. 13 (right) shows the degree of suppression afforded by in-channel cancellation (yellow: no suppression, blue: strong suppression). Clear suppression is evident in a small number of time-frequency pixels, which luckily correspond to pixels of high

amplitude in the excitation pattern (left). Within certain pixels the attenuation reaches 20 dB or more, which might make it easier to hear features of a weak target (e.g. other voice) that happen to fall within those pixels. Whether this can translate into a benefit in terms of intelligibility, for example as an explanation of the "cocktail party effect", is beyond the scope of this paper.

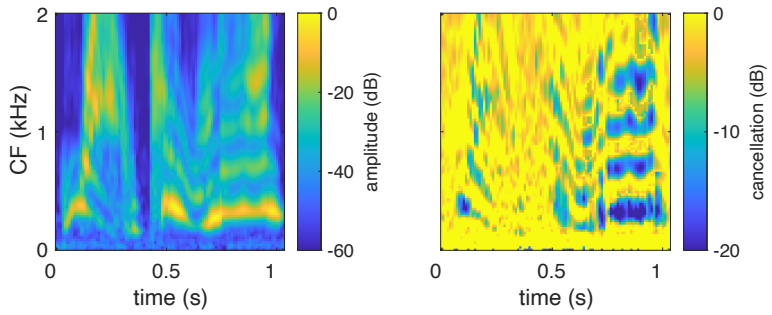

Fig. 13. Left: spectro-temporal excitation pattern in response to speech snippet "Wow Cool!". Right: attenuation provided by cancellation filter (ratio of output power to input power per channel and time step).

Pitch, pitch change, tonality, transparency

A by-product of cancellation is the automatically-determined delay parameter T. The original harmonic cancellation model produced a single estimate that could serve as a cue to the pitch of a periodic sound (de Cheveigné 1998). The in-channel model differs in that it offers *multiple* delay estimates T_k, each corresponding to the period of a partial, or of a group of harmonically-related partials that dominates a channel. These can be variously interpreted as cues to a pitch local to a spectral region (accessible to conscience via attention to that spectral region), or as a set of "partial pitches" with perceptual reality at a sub-attentive level (analogous to the "spectral pitches" invoked by Terhardt 1974). Certain sounds that lack a clear period, such as inharmonic tones, or bandpass noise, may also be judged to have a pitch-like quality (sometimes referred to as "tonality") that the in-channel cancellation might help explain. Speculating, a condition for perceived tonality may be the existence of peripheral channels for which local cancellation is approximately successful.

Demany and Ramos (2005, Demany et al 2011) found that subjects could detect the direction of *pitch change* for one partial within a non-harmonic 5-partial complex, despite the fact that that partial's pitch was not perceptually accessible before the change. McPherson et al (2018) likewise found that subjects could judge the direction of pitch change between

inharmonic complexes (at least for short delays between sounds). These phenomena suggest the existence of a battery of "frequency shift detectors" as reviewed by Demany and Semal (2018). The in-channel cancellation model might help explain how such shift detectors are implemented. The period T_k within a channel can be estimated by scanning the outputs of an array of cancellation filters for a minimum, as schematized in Fig. 14 for one channel. The direction of period *change* can then be found by observing the filter output for shorter delays (left of the original minimum): a decrease indicates an upward frequency shift, an increase indicates a downward shift. The ensemble of direction estimates (upward vs downward) can be summarized across channels by an averaging or voting mechanism (e.g. to explain McPherson et al 2018, or Allik et al 1989), or by selectively attending to those channels for which there is change (e.g. to explain Demany and Ramos 2005). This account is applicable to stimuli that lack an overall unitary periodicity (e.g. an inharmonic complex), in contrast to the original harmonic cancellation model that assumes the same period across channels. The in-channel cancellation model is an improvement in this respect.

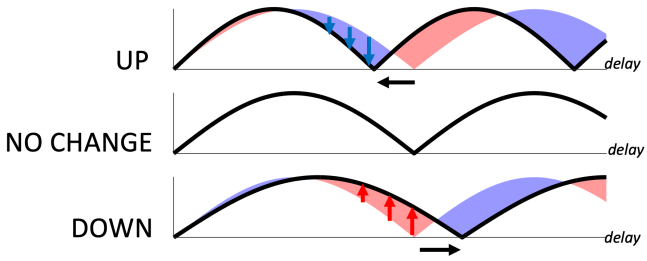

Fig. 14. Schematic principle of an in-channel frequency shift detector. Each plot represents the amplitude at the output of a cancellation filter as a function of its delay parameter (or across an array of such filters indexed by delay) at two moments in time. For an upward change (top), the amplitude drops to the left of the original minimum (shorter delay), for a downward change (longer delay) it increases. This detector is instantiated independently in all channels.

The availability of a mechanism that can reduce the masking power of certain sounds implies for those sounds a property of *transparency*, that may be beneficial in a musical context that involves building a complex sound structure with multiple elements (polyphony). The same property may be detrimental, for example for perceptual coding of audio signals, as it enhances the detectability of quantization noise (Johnston 1988).

How do in-channel cancellation and harmonic cancellation models relate to each other?

The two models are similar in principle and structure. In-channel cancellation might seem more complex because it involves more parameters (the value of T_k in each channel), and thus less parsimonious. However, the T_k are chosen automatically and thus do not constitute "free" parameters (they cannot be tweaked to improve the fit to a data set). One might argue instead that the in-channel model is *simpler* because it is implemented locally within each channel, with no need for a cross-channel mechanism to derive a common T across channels and distribute that estimate to each channel.

The two models are equally able to cancel a perfectly harmonic background (same period in all frequency regions), but the in-channel model can additionally attenuate backgrounds that are locally harmonic, as well as those that are spectrally sparse (e.g. Roberts and Brunstrom 2001; Deroche et al 2014; Popham et al 2018; Steinmetzger and Rosen 2023). Whereas harmonic cancellation in its standard formulation would be confused by the lack of a clear "period" for such maskers, in-channel cancellation remains effective, either because the inharmonic spectrum may be approximated by a harmonic series over a limited frequency range, or because partials can be cancelled individually, or both. The relationship is analogous to that between the original EC model of binaural unmasking (Durlach 1963) and modified EC models (Culling and Summerfield 1994; Breebart and Kohlrausch 2001; Ackeroyd 2004).

The possibility of implementing in-channel cancellation within tonotopically differentiated neural tissue may make it physiologically more plausible than harmonic cancellation as all operations are tonotopically local and require no cross-channel communication.

Invariance, unconscious inference, and predictive coding

One can argue that a major goal of auditory processing is to ensure invariance, both to *irrelevant sound dimensions* for the purpose of classification, and to *competing sound sources* for the purpose of auditory scene analysis. Cochlear frequency resolution or temporal resolution can be interpreted in this light, and binaural, harmonic, or in-channel cancellation as complementing them. For example, cochlear filtering is effective to ensure invariance of a narrowband target to the presence of off-band energy (Fig. 2 top), and in-channel cancellation extends this ability to an on-frequency tonal masker for which cochlear filtering does not suffice (Fig. 2 bottom).

Cochlear filtering and cancellation both contribute to invariance by suppressing the masker, but this operation may affect the target: target features that fall within discarded channels or on zeros of the cancellation filter (Fig. 3, right) are lost. However, incomplete sensory patterns can be accommodated via a process of *unconscience inference* (Helmholtz 1867) by which an internal model of a perceptual object is constrained by the available (albeit incomplete) information (de Cheveigné 2021).

In-channel cancellation can also be seen as a form of *predictive coding* (Barlow and Rosenblith 1961; Friston 2018) operating at the very first stage of post-cochlear processing. The outcome of this coding is both a delay parameter (and possibly a goodness of fit measure) for each channel, that contribute to characterize the stimulus, and an "error signal" that may reveal a weaker concurrent target. In-channel cancellation is related to linear predictive coding (LPC) (Atal 2006), and one might speculate that other forms of LPC could be approximated by time domain processing within the auditory brainstem.

Neural implementation

Two issues must be addressed: How can the *linear filtering* operations required by the model be approximated with neurons? How can they be effective given the constraints of *non-linear peripheral transduction* and *stochastic neural coding*? A simple neural approximation of the cancellation filter of Fig. 3 is described by de Cheveigné (2021) who also reviewed potential sites where such a circuit could be implemented within the brain. In brief, there are multiple sites within the auditory brainstem where the necessary excitatory-inhibitory interactions could occur, from dendritic fields within the cochlear nucleus to dendritic fields within the inferior colliculus. In such a circuit, cancellation would operate on *instantaneous spike probabilities*, and one might wonder whether this allows sufficient linearity and dynamic range for effective filtering. To settle this issue requires simulation with realistic models of transduction and neural processing, which is beyond the scope of this paper, however previous modelling demonstrated effective processing for synthetic and speech stimuli using recorded auditory-nerve patterns (Palmer 1990) and simulated spike trains (de Cheveigné 1993, Guest and Oxenham 2019). Evidence for spike-based subtractive processing as required by the EC model has been reported in the auditory brainstsem, e.g. Franken et al (2021). The neural cancellation filter has the same structure as the same-

frequency inhibitory-excitatory (SFIE) circuit that Nelson and Carney (2004) proposed to explain amplitude-modulation tuning in the inferior colliculus.

Spectro-temporal fan-out in the brainstem

A remarkable feature of the auditory system is the progressive "fan-out" from the single channel of acoustic vibration at each ear to ~3000 inner hair cells, ~30000 auditory nerve fibers, ~100000 neurons in the cochlear nucleus (Hinojosa 2011), and millions of cells within the auditory cortex. Response properties are usually more complex at higher levels. In-channel cancellation could participate in such a fan-out by augmenting each peripheral channel with an additional cancellation-filtered channel. Multiple, diverse transforms help pattern recognition algorithms ensure invariance to irrelevant feature dimensions (Duda et al 2012) or, as we suggest here, to background sounds. Random transforms are sufficient for pattern recognition as long as they include convolution and non-linearity and are sufficiently numerous and diverse (e.g. Gauthier 2021), but there is a benefit to selecting a priori useful transforms. In-channel cancellation ensures invariance to a class of interfering stimuli (harmonic and/or spectrally sparse), which argues for including this transform within the bouquet. In-channel cancellation is arguably a "good thing to have" for the auditory brainstem.

A classic view of auditory processing is that cochlear filter outputs are demodulated and represented as a slowly-varying spike rate, in which case spectral resolution is determined entirely by that of the cochlea. Neural time-domain processing offers a wider range of possibilities. Using a sampled-signal notation for convenience, the cochlear filter bank can be approximated as a M-column matrix $\boldsymbol{M}$ of finite impulse responses of order N. Filtering amounts to multiplying a N-column matrix $\boldsymbol{X}$ of delayed stimulus signals (delays $1 \cdots N$) by the matrix $\boldsymbol{M}$ to obtain the matrix $\boldsymbol{Y} = \boldsymbol{XM}$ of cochlear-filtered signals (one channel per column). If the matrix $\boldsymbol{M}$ allows an *inverse*, (i.e. if its rank is larger than N), multiplying $\boldsymbol{Y}$ by this inverse would reconstruct $\boldsymbol{X} = \boldsymbol{YM}^{-1}$, i.e. the acoustic waveform (a single column of $\boldsymbol{X}$) could be reconstructed as a weighted sum of cochlear filter outputs. Furthermore, supposing that a task can be addressed by applying some filter of impulse response $\boldsymbol{h}$ to the acoustic waveform. The output of that filter is available by multiplying $\boldsymbol{Y}$ by $\boldsymbol{M}^{-1}\boldsymbol{h}$. In other words, *any filter* (of order at most N) applicable to the acoustic stimulus can be implemented within the auditory system by taking a *weighted sum of cochlear filter outputs*. The operations applied

to $\boldsymbol{Y}$ (cochlear filter bank outputs) are purely scalar: no delays are required, although the availability of delays might ease the implementation.

This rosy perspective should be tempered because it assumes perfect linearity, whereas we know that transduction and neural processing have imperfect linearity and limited dynamic range. Measurements of binaural or harmonicity-based unmasking (reviewed in de Cheveigné 2021, p10) suggest that the benefit of post-cochlear filtering might be limited to about 3-15 dB. Cochlear filtering is required for greater attenuations, and thus the concepts of "critical band" and "resolvability" remain entirely relevant.

By drawing on both cochlear filtering and time-domain neural processing, in-channel cancellation resembles other hybrid models such the lateral inhibitory network (LIN) of Shamma (1985) or the phase opponency of Carney et al (2002), or earlier proposals of a "second filter" (Huggins and Licklider 1951) or the more recent idea of "synthetic delay lines" (de Cheveigné and Pressnitzer 2006). The interplay between cochlear filtering and time-domain processing is integral to all these models.

Variants of the cancellation filter

A partial of frequency f can be cancelled with a delay equal to any multiple $T = k/f$ of its period. There might be an advantage of choosing a larger multiple, as a cancellation filter tuned to a higher value of k imposes less attenuation to the spectral region immediately adjacent the peak of the peripheral filter, as illustrated in Fig. 15.

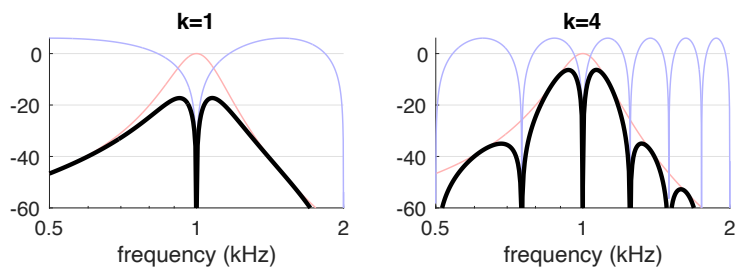

Fig. 15. Transfer functions of compound filter with CF=1 kHz for $T = 1$ ms (left) or $T = 4$ ms (right). Both offer perfect rejection at 1 kHz, but the filter on the right imposes less attenuation on the spectral region near 1 kHz.

On the other hand, it has been argued that there may be a penalty for longer delays in the auditory system (as reviewed by de Cheveigné and Pressnitzer 2006), in which case $k = 1$

might be preferable. Longer delays are still required to cancel multiple harmonically-related partials within a channel, but they might be avoided by applying instead a cascade of two cancellation filters tuned to the two most prominent harmonics within that channel. As illustrated in Fig. 2 for a 200 Hz harmonic complex, in channels up to about 1 kHz essentially all the power is accounted for by just two partials, which a two-cancellation-filter cascade could suppress. These variants are worth considering when simulating the model or searching for its correlates within the brain.

Conclusion

This paper considered the hypothesis of a simple filtering mechanism implemented at the output of the cochlea for the purpose of ensuring invariance with respect to maskers that are harmonic, locally harmonic, or spectrally sparse. The model accounts qualitatively for the masking asymmetry between pure tone and narrowband noise, and between harmonic and inharmonic complex tones. The filter involves a channel-specific delay parameter that is estimated automatically within each channel. As a spinoff of this unmasking mechanism, the delay parameter might also serve as a cue to pitch, or to detect frequency change between inharmonic stimuli that do not evoke a clear pitch. The model was discussed as an example of a wider class of putative processing mechanisms that associate quasi-linear filtering in the cochlea with time-domain neural processing in the brainstem.

Acknowledgments

This work was supported by grants ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL, and ANR-17-EURE-0017.

References

Akeroyd MA (2004) The across frequency independence of equalization of interaural time delay in the equalization-cancellation model of binaural unmasking. J. Acoust. Soc. Am. 116:1135–1148.

Allik, J., E. N. Dzhafarov, A. J. M. Houtsma, J. Ross, and N. J. Versfeld. 1989. 'Pitch Motion with Random Chord Sequences'. *Perception & Psychophysics* 46 (6): 513–27. https://doi.org/10.3758/BF03208148.

B.S. Atal (2006). "The history of linear prediction". IEEE Signal Processing Magazine. 23 (2): 154–161.

Barlow, H., B. & Rosenblith, W., A. Possible principles underlying the transformations of sensory messages, pages 217-234. MIT Press, 1961

Breebaart J, van de Par S, Kohlrausch A (2001) Binaural processing model based on contralateral inhibition. I. Model structure. The Journal of the Acoustical Society of America 110:1074–1088.

Carney, L. H., Heinz, M. G., Evilsizer, M. E., Gilkey, R. H., and Colburn, H. S. (2002) Auditory Phase Opponency: A Temporal Model for Masked Detection at Low Frequencies. Acta Acustica united with Acustica, 88:334-347

Culling JF, Summerfield Q (1994) Binaural segregation of concurrent sounds involves within-channel rather than across-channel processes. The Journal of the Acoustical Society of America 95:2915–2915.

Demany L, Ramos C (2005) On the binding of successive sounds: Perceiving shifts in nonperceived pitches. J Acoust Soc Am 117:833–841.

Demany, Laurent, Catherine Semal, and Daniel Pressnitzer. 2011. 'Implicit versus Explicit Frequency Comparisons: Two Mechanisms of Auditory Change Detection.' *Journal of Experimental Psychology: Human Perception and Performance* 37 (2): 597–605.

Deroche, M. L. D., Culling, J. F., Chatterjee, M., and Limb, C. J. (2014). "Speech recognition against harmonic and inharmonic complexes: Spectral dips and periodicity," J. Acoust. Soc. Am. 135, 2873–2884.

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

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

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

de Cheveigné, A. and Pressnitzer, D. (2006) The case of the missing delay lines: synthetic delays obtained by cross-channel phase interaction J. Acoust. Soc. Am., 119, 3908-3918.

de Cheveigné, A (2021). Harmonic Cancellation - A Fundamental of Auditory Scene Analysis. Trends in Hearing, 25, https://doi.org/10.1177/23312165211041422.

Demany, Laurent, and Christophe Ramos. 2005. On the Binding of Successive Sounds: Perceiving Shifts in Nonperceived Pitches. *J. Acoust. Soc. Am.* 117 (2): 9.

Demany, Laurent, Catherine Semal, and Daniel Pressnitzer. 2011. 'Implicit versus Explicit Frequency Comparisons: Two Mechanisms of Auditory Change Detection.' *Journal of Experimental Psychology: Human Perception and Performance* 37 (2): 597–605. https://doi.org/10.1037/a0020368.

Demany, Laurent, and Catherine Semal. 2018. 'Automatic Frequency-Shift Detection in the Auditory System: A Review of Psychophysical Findings'. *Neuroscience* 389 (October): 30–40. https://doi.org/10.1016/j.neuroscience.2017.08.045.

Duda R. O., Hart P. E., & Stork D. G. (2012). Pattern classification. John Wiley & Sons.

FLETCHER H. Auditory patterns. Rev. Mod. Phys. 12:47–65, 1940.

Franken T. P., Bondy B. J., Haimes D. B., Goldwyn J. H., Golding N. L., Smith P. H., & Joris P. X. (2021). Glycinergic axonal inhibition subserves acute spatial sensitivity to sudden increases in sound intensity. eLife, 10, e62183. https://elifesciences.org/articles/62183 https://doi.org/10.7554/eLife.62183

Friston K. (2018). Does predictive coding have a future? Nature Neuroscience, 21(8), 1019–1021. http://www.nature.com/articles/s41593-018-0200-7 https://doi.org/10.1038/s41593-018-0200-7.

Gauthier, Daniel J., Erik Bollt, Aaron Griffith, and Wendson A. S. Barbosa. 2021. 'Next Generation Reservoir Computing'. *Nature Communications* 12 (1): 5564. https://doi.org/10.1038/s41467-021-25801-2.

Guest D. R., & Oxenham A. J. (2019). The role of pitch and harmonic cancellation when listening to speech in harmonic background sounds. Journal of the Acoustical Society of America, 145(5), 3011–3023. http://asa.scitation.org/doi/10.1121/1.5102169 https://doi.org/10.1121/1.5102169

Helman, R. (1972) Asymmetry of masking between noise and tone. Perception and psychophysics 11, 241-246.

Helmholtz H. (1867). Handbuch der Physiologischen Optik (English tranl.: 1924 JPC Southall as Treatise on Physiological Optics). Voss.

Hinojosa, R., & Nelson, E. G. (2011). Cochlear nucleus neuron analysis in individuals with presbycusis: Cochlear Nucleus Neuron Analysis. *The Laryngoscope*, *121*(12), 2641–2648. https://doi.org/10.1002/lary.22383

Johnston, J.D. 1988. 'Transform Coding of Audio Signals Using Perceptual Noise Criteria'. *IEEE Journal on Selected Areas in Communications* 6 (2): 314–23. https://doi.org/10.1109/49.608.

Kim, D. O., Carney, L., & Kuwada, S. (2020). Amplitude modulation transfer functions reveal opposing populations within both the inferior colliculus and medial geniculate body. *Journal of Neurophysiology*, *124*(4), 1198–1215. https://doi.org/10.1152/jn.00279.2020

Licklider J. C. R. (1951). A duplex theory of pitch perception. Experientia, 7, 128–134. https://doi.org/10.1007/BF02156143

Hall, J. L. (1997). Asymmetry of masking revisited: Generalization of masker and probe bandwidth. J.Acoust. Soc. Am., 101:1023{1033.

Holdsworth, J., Nimmo-Smith, I., Patterson, R. D., and Rice, P. (1988) Implementing a GammaTone filter bank'' SVOS final report, annex C, MRC Applied Psychology Unit Tech. Rep. (unpublished).

Huggins W., & Licklider J. (1951). Place mechanisms of auditory frequency analysis. Journal of the Acoustical Society of America, 23, 290–299, https://doi.org/10.1121/1.1906760

McPherson, Malinda J., and Josh H. McDermott. 2018. 'Diversity in Pitch Perception Revealed by Task Dependence'. *Nature Human Behaviour* 2 (1): 52–66.

Moore, B. C. J., and Glasberg, B. R. (1983) Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, J. Acoust. Soc. Am. 74, 750–753.

Moore, B. C. J. (1995). Frequency analysis and masking. In B. C. J. Moore (Ed.), *Hearing* (pp. 161–205). Academic Press.

Moore, Brian C. J., Joseph I. Alcántara, and Torsten Dau. 1998. 'Masking Patterns for Sinusoidal and Narrow-Band Noise Maskers'. *The Journal of the Acoustical Society of America* 104 (2): 1023–38. https://doi.org/10.1121/1.423321.

Moore, Brian C.J., and Hedwig E. Gockel. 2011. 'Resolvability of Components in Complex Tones and Implications for Theories of Pitch Perception'. *Hearing Research* 276 (1–2): 88–97. https://doi.org/10.1016/j.heares.2011.01.003.

Nelson, P. C., & Carney, L. H. (2004). A phenomenological model of peripheral and central neural responses to amplitude-modulated tones. *The Journal of the Acoustical Society of America*, *116*(4), 2173–2186. https://doi.org/10.1121/1.1784442

Palmer, A. R. 1990. ‘‘The representation of the spectra and fundamental frequencies of steady-state single- and double-vowel sounds in the temporal discharge patterns of guinea pig cochlear-nerve fibers,’’ J. Acoust. Soc. Am. 88, 1412–1426.

Popham S., Boebinger D., Ellis D. P. W.,Kawahara H.,&McDermott J. H. (2018). Inharmonic speech reveals the role of harmonicity in the cocktail party problem. Nature Communications, 9(1), 21–22. http://www.nature.com/articles/s41467-018-04551-8 https://doi.org/10.1038/s41467-018-04551-8.

Roberts B, Brunstrom JM (2001) Perceptual fusion and fragmentation of complex tones made inharmonic by applying different degrees of frequency shift and spectral stretch. The Journal of the Acoustical Society of America 110:2479–2490.

S. A. Shamma: Speech processing in the auditory system-II: Lateral inhibition and the central processing of speech-evoked activity in the auditory nerve. J. Acoust. Soc. Am. 78 (1985) 1622–1632

Siedenburg, Kai, Jackson Graves, and Daniel Pressnitzer. 2022. ‘A Unitary Model of Auditory Frequency Change Perception’. *BioRxiv*, January, 2022.06.16.496520.

Slaney M. (1993). An efficient implementation of the Patterson- auditory filter bank (technical report No. 35). Apple Computer.

Smoorenburg, G.F. and Coninx, F. (1980) Masking of short probe sounds by tone bursts with a sweeping frequency. Hearing Research 3, 301-316.

Srulovicz, P., and Goldstein, J. L. (1983). "A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum," J. Acoust. Soc. Am. 73, 1266-1276.

Steinmetzger, K., Rosen, S. (2023) No evidence for a benefit of masker harmonicity in the perception of speech in noise, J. Acoust. Soc. Am. 153, 1064, https://doi.org/10.1121/10.0017065.

Terhardt E. (1974). Pitch, consonance, and harmony. Journal of the Acoustical Society of America, 55(5), 1061–1069. http://asa.scitation.org/doi/10.1121/1.1914648 https://doi.org/10.1121/1.1914648.

Verhey, J. L. (2002). Modeling the influence of inherent envelope fluctuations in simultaneous masking experiments. J. Acoust. Soc. Am., 111:1018-1025.