version 20221127

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. The delay is tuned automatically with a criterion of minimum power, independently for each channel. 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 channel is thus split into two subchannels, one filtered and the other not, either of which can be attended to depending on the task. Here, the model is used to explain the masking asymmetry between pure tones and noise. According to the model, a noise target masked by a tone is more easily detectable, thanks to cancellation, than a tone target masked by noise, as indeed is reported in the literature. The in-channel cancellation model is one of a wider class of models, monaural or binaural, that apply cancellation to suppress 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 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, 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). A pure tone and a narrowband noise elicit roughly the same

distribution of power across the basilar membrane, so we should expect their masking effects to be the same if power were the sole criterion. This departure from the power spectrum model suggests the operation of an unmasking mechanism applicable to a pure tone masker (Hall 1997).

A similar masking asymmetry has been observed between harmonic and inharmonic (or noise-like) probes and 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 the operation of an unmasking mechanism specific to harmonic maskers. This is analogous to spatial unmasking, where it appears that 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 (as from a spatially diffuse sound field).

Binaural unmasking is well accounted for by Durlach's (1963) equalization-cancellation (EC) model and its variants, that assume that neural signals from both ears are differentially scaled and delayed (equalization), and then subtracted centrally (cancellation). Durlach's (1963) model assumes, tacitly, that the same parameters (delay and scale factor) are applied to every peripheral channel, and thus it cannot explain the degree of unmasking that is observed with maskers crafted to have different binaural properties (e.g. different interaural delay) in different frequency bands. To account for those observations, more flexible “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).

An analogous model was proposed to explain harmonic unmasking (de Cheveigné 1993, 2021). The “harmonic cancellation” model assumes that a neural signal is delayed and subtracted from the non-delayed signal, with a delay equal to the period of the interfering sound. This effectively suppresses correlates of a periodic (harmonic) masker. The harmonic cancellation model assumes that the same delay is applied to every peripheral channel, and thus it cannot explain unmasking observed with inharmonic stimuli obtained by stretching, shifting, or jittering the partials of a harmonic complex (as reviewed in de Cheveigné 2021). The model proposed here relaxes the single-delay assumption, taking inspiration from the modified EC model, by allowing for a different delay in each channel. However, the in-channel cancellation

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

An alternative explanation for the tone/noise masking asymmetry suggests 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 by that explanation.

The hypothesis of automatically-tuned cancellation filters within *every* peripheral channel opens some interesting and potentially far-reaching perspectives. First, it highlights the role of time-domain signal processing in the auditory brainstem, with the implication 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 fundamental goal of auditory processing, and cancellation as a fundamental operation to achieve that goal. Cancellation does not guarantee the integrity of the representation of a target sound, and thus it needs to work hand-in-hand with a Helmholtzian process that fits incomplete or distorted representations to an internal model. 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, as a by-product, a single period estimate (a potential cue to pitch, see de Cheveigné 1998), the in-channel version offers multiple local estimates that may be useful to explain sensitivity to *pitch change* of stimuli that lack an overall pitch.

Methods

The model

A simplified linear model of cochlear filtering is used to derive qualitative predictions to be compared with experimental results. For simplicity and clarity, effects of non-linear

transduction, stochastic neural coding, compression, and so on, are not considered. The initial stage of auditory processing is modeled as a linear filter bank, followed by a delay-and-subtract cancellation filter. The auditory brain is assumed to have access to both the original signal, and the cancellation filter output. The delay is estimated automatically and independently in each channel, and this estimate is also available to the brain. 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 at the lowest CFs. Each channel attenuates all but a narrow frequency region, implying that its output is relatively insensitive to the presence signal features outside that region. However, attenuation is not infinite at any frequency.

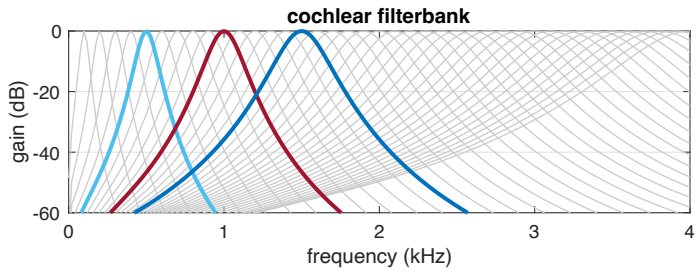

Fig. 1. Transfer functions of selected channels of the gammatone filter bank. Channels with CF=0.5, 1 and 2 kHz are highlighted.

The selectivity of cochlear filters can be exploited to “protect” a target sound from a competing masker. This is illustrated in Fig. 2 (top) for a 1 kHz probe in a wideband masker (equal RMS). Channels near 1 kHz are purely dominated by the probe. As another example, individual partials of a 200 Hz harmonic complex tone can be isolated within channels with CFs near each partial's frequency (Fig. 2, middle). Schematically, two ranges of CFs can be distinguished in this example. Below 1 kHz (5th harmonic), isolation appears perfect within a

subset of channels close to each partial's frequency, with intermediate channels responding to at most two neighboring partials. Above 1 kHz, isolation is less perfect (the peaks of the red curve do not reach one), and channels tend to respond to more than two partials (dotted line).

This parallels the classic distinction between “resolvable” and “unresolvable” partials of a complex (Moore and Gockel 2011). Resolvability is often attributed to the salience of ripples in the excitation pattern (e.g. Fig. 1 in Moore and Gockel 2001), but Fig. 2 (middle) suggests an alternative account: a partial is resolvable if it dominates at least one peripheral channel (i.e. its power relative to other stimulus components is above 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).

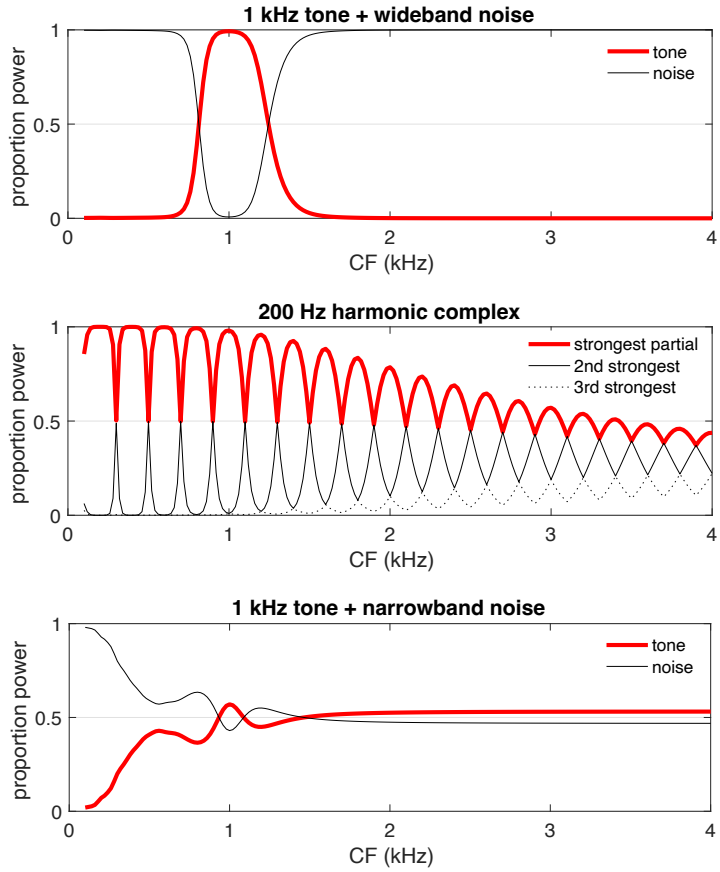

Fig. 2. Proportion of power at the output of a cochlear filterbank attributable to different parts of the stimulus. Top: mixture of a pure tone and wideband noise with equal RMS amplitude (over a 20 kHz bandwidth). Red, black: proportion of power attributable to the

pure tone and the noise, respectively. 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: same as top but the noise is narrowband, centered on 1 kHz with a 1 ERB bandwidth.

Both examples illustrate the usefulness of peripheral filtering for the purpose of hearing out 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), with equal RMS amplitudes. No channel is clearly dominated by either the pure tone or the narrowband noise (except in the lowest channels, for which both are in any case severely attenuated, c.f. Fig. 1). Based on this simulation one would expect high thresholds for detecting a tone probe in a narrowband noise or vice-versa. Furthermore, one would expect thresholds to be *similarly high* in both configurations.

The cancellation filter is 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.

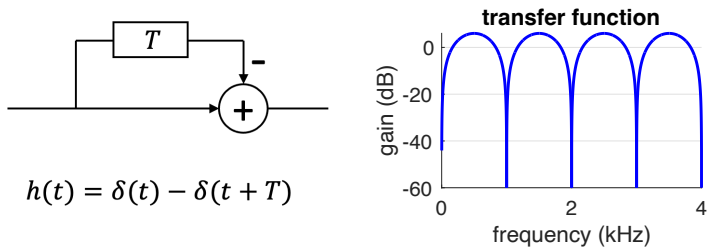

Fig.3. Left: schema and impulse response of a cancellation filter. Right: transfer function for T=1 ms.

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 ingredients, in particular it has *infinite* attenuation at 1 kHz, and reduced gain at frequencies remote from its CF. The output of this compound filter is invariant to the presence, or absence, of a component at 1 kHz.

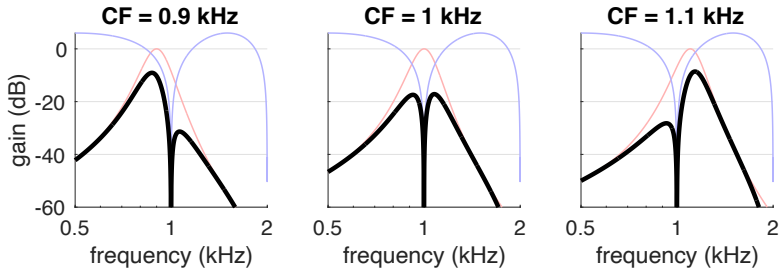

Fig. 4. Transfer function of the cascade of a gammatone filter and a cancellation filter with a delay parameter T=1 ms for three CFs (as indicated). The thin colored lines are the transfer functions of the component filters (red: gammatone, blue: cancellation).

Finding T

The in-channel cancellation model makes an important assumption: the delay T is determined automatically within each channel, based on the signal within that channel. This is achieved by searching for the non-zero value of T that minimizes the power at the output of the cancellation filter, relative to its input (Fig. 5). 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).

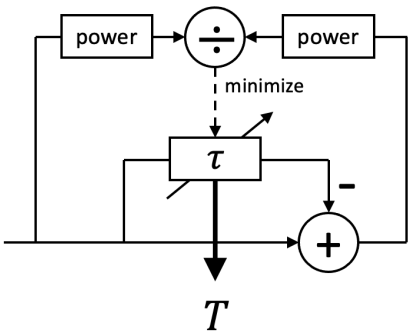

Fig. 5. Automatic determination of the delay T. A range of potential values is searched for the smallest ratio of cancellation filter output power to input power.

A couple of details need elaborating. First, the power ratio must be estimated over a window long enough to get a stable estimate, yet short enough to follow changes in a non-stationary signal. In the simulations the window size was set to 30 ms. Second, for a purely periodic signal (e.g. a pure tone), the ratio is zero for the period and *all its multiples*, implying an infinite number of equally valid candidates for T. For definiteness, it may be convenient to choose among candidates the first for which the power ratio is below some threshold θ, e.g. $\theta = 0.1$ (see de Cheveigné and Kawahara 2002 for the rationale and more details). This parameter is not critical.

Figure 6 plots the inverse of the automatically-determined delay T within each channel as a function of CF for three stimuli: a 1 kHz pure tone (left), wideband noise (center), and narrowband noise (right). For the pure tone, the inverse estimate is $1/T$ = 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 CF and stabilizes near 1 kHz at high CFs. These examples illustrate how the delay T is automatically chosen for three common stimuli. Additional examples are provided later on.

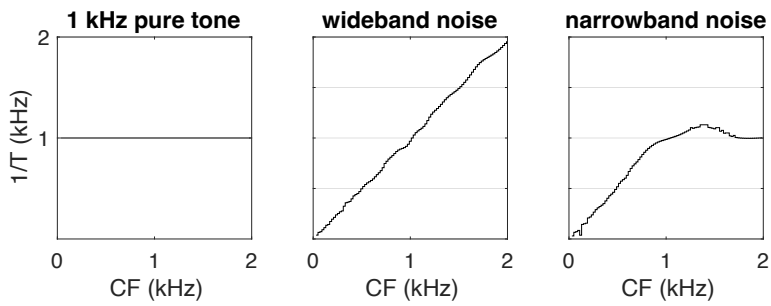

Fig. 6. Inverse of estimated delay T for each channel as a function of its CF for two stimuli. For the pure tone (left), all channels yield $1/T = 1\ kHz$. For white noise (center), the value is approximately $1/T$ =CF. For narrowband noise (right) it follows $1/T$ =CF at low frequencies and is close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

For each peripheral channel, the auditory system is assumed to have access to both its cancellation-filtered and unfiltered output, with the ability to *attend* to either depending on

the task. It is also assumed to have access to the automatically-estimated delay parameters T for each channel. Since they are automatically determined, they do not constitute free parameters in the model.

Results

Response to pure tone and narrowband noise

For a pure tone, assuming the delay T 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).

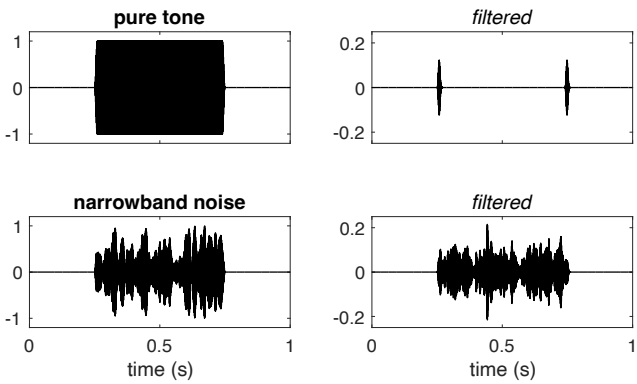

Fig. 7. Effect of the compound filter within the CF=1 kHz channel. Left: stimulus, right: compound filter output. Top: 1 kHz pure tone, bottom: narrowband noise centered on 1 kHz (width 0.5 ERB ≈ 60 Hz) and wideband noise. Stimuli are 500 ms in duration and shaped with 10 ms raised-cosine onset and offset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

The same is not true for narrowband noise (Fig. 7 bottom right). For that stimulus, the compound filter fails to cancel any portion. In both examples, the delay T was chosen automatically as described earlier.

Figure 8 shows the spectro-temporal excitation pattern across peripheral filter channels as a color-coded image, for the pure tone (top) and narrowband noise (bottom), 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. For the pure tone, cancellation suppresses the response for most of the duration of the stimulus (top right), but the same is not true for narrowband noise (bottom right). 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.

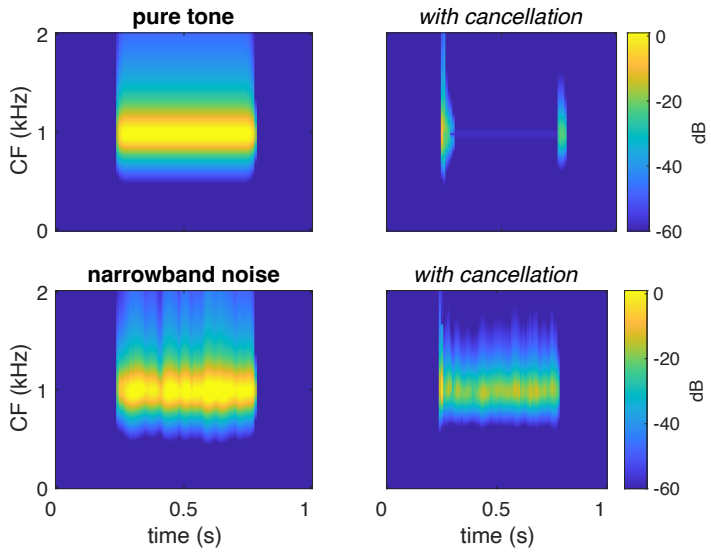

Fig. 8. Spectro-temporal excitation patterns before (left) and after (right) cancellation. For the pure tone (top), after cancellation the interval between onset and offset is almost perfectly quiescent, in contrast to the narrowband noise (bottom). The auditory brain is assumed to have access to both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

Pure tone/noise masking asymmetry

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).

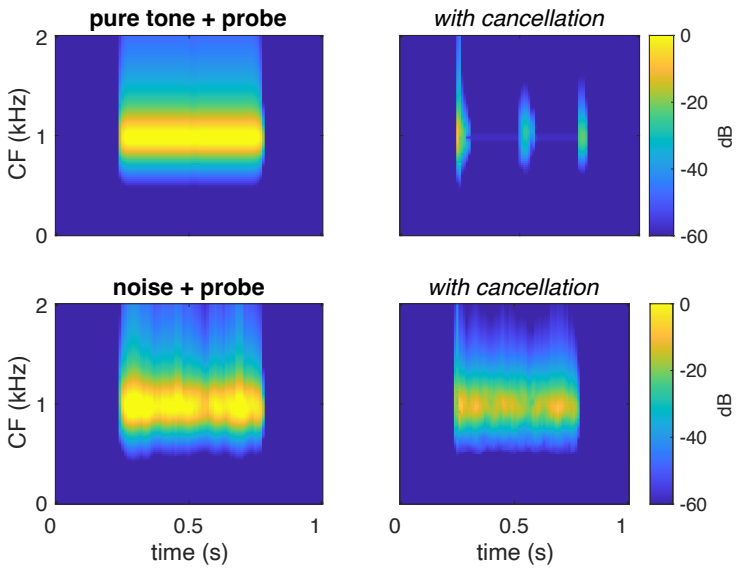

Fig. 9. Spectro-temporal excitation patterns before (left) and after (right) cancellation for a probe added to a masker. The masker was a 1 kHz tone (top) or a narrowband noise of width 0.5 ERB (bottom). The probe was a 30 ms pulse of narrowband noise temporally centered within the masker, with an amplitude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

This is qualitatively consistent with the observations of Helman (1972) and others: 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 principal result of this paper. The rest of this section examines a few other examples of interest.

Smoorenburg's sweep tone masker

Smoorenburg and Coninx (1980) found that a pure-tone was up to 20 dB less potent than a sweep tone at masking a short pure tone probe with a frequency that matched that of the pure tone, or the instantaneous frequency of the sweep at the probe's position. The standard power-spectrum model of masking predicts *less* masking for the sweep because the sweep 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 narrow-band probe temporally centered on a 1 kHz pure tone masker (left) or a sweep-tone masker (right). The latter swept logarithmically from 500 Hz to 2 kHz in 0.5 s (sweep rate 4

octaves/s). The probe was a 30 ms burst of narrowband noise (0.5 ERB bandwidth) centered at 1 kHz, temporally positioned to coincide with the instantaneous frequency of the masker, with an amplitude equal to that of the masker. 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

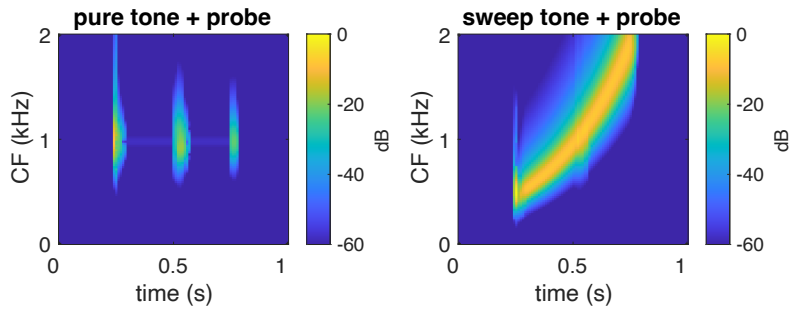

Fig. 10. Spectro-temporal excitation patterns before (left) and after (right) cancellation for a probe added to a sweep-tone masker. The probe is the same as in Fig. 9. Cancellation allows the probe to emerge (visually) from the pure tone but not the sweep.

Harmonic and inharmonic maskers

Figure 11 shows the inverse delay parameter $1/T$ 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, for 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, $1/T$ appears to follow either the frequency of an individual partial, or what seems to be a common subharmonic of neighbouring partials, and for higher CFs, it 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 tends to follow the fundamental period of the harmonic tone, or a "local fundamental" of the inharmonic tone, rather than the period of individual partials. This has little effect on subsequent results.

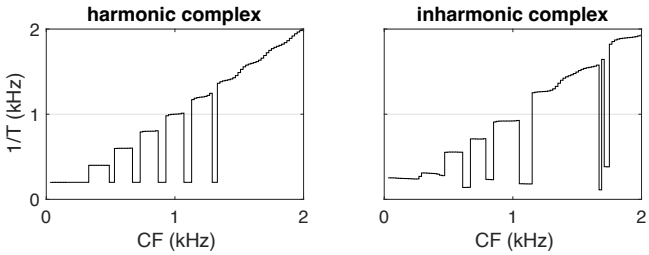

Fig. 11. Inverse of the delay parameter estimated within each channel as a function of its CF, for a harmonic complex (left) or an inharmonic complex (right).

Figure 12 shows spectro-temporal excitation patterns before (left) and after (right) cancellation for the 200 Hz harmonic complex tone (top), and the inharmonic complex (bottom). For the harmonic complex, cancellation effectively suppresses the excitation pattern over most of the duration of the stimulus (bottom left).

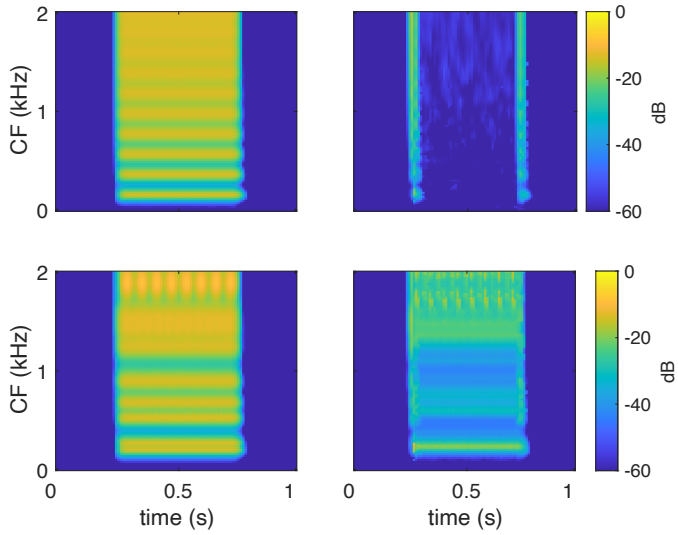

Fig. 12. Spectro-temporal excitation patterns before (top) and after (bottom) cancellation for four different stimuli as labeled (see text). Cancellation is most effective for the harmonic complex (left) and least effective for the noise-like “dense” inharmonic complex (right).

For the inharmonic complex (bottom center), the effectiveness of cancellation is reduced but nonetheless greater than for a noise-like stimulus (not shown). The latter may explain why inharmonic maskers share some of the properties of harmonic maskers, being less deleterious than noise-like maskers of similar spectral shape, albeit more than harmonic maskers.

As a final example, Fig. 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). The benefit is limited to certain time-frequency pixels, which luckily correspond to pixels of high amplitude in the excitation pattern (left). Within certain of these pixels, the attenuation reaches 20 dB or more. The purpose of this example is to show that in-channel cancellation can operate, in principle, with real-world sounds such as speech. 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.

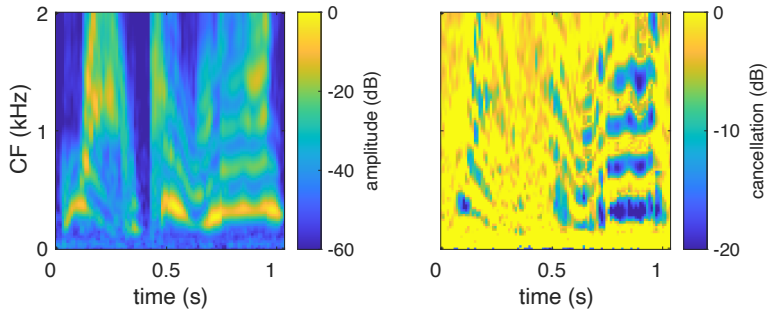

Fig. 13. Left: spectro-temporal excitation pattern in response to speech snippet "Wow Cool!". Right: attenuation provided by cancellation filter (ratio of output power to input power per channel and time step).

Discussion

The in-channel cancellation model offers a simple explanation for the asymmetry of masking between tone and noise. It complements 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 are qualitative, based on a simplified linear approximation of cochlear and neural signal processing. Whether the qualitative trends hold with more realistic non-linear transduction and stochastic neural processing, and whether they can be coaxed into a quantitative fit with behavioural results remains to be determined. This is food for future studies. This section discusses some limits of the model, and speculates on its possible significance as an example

of a wider class of sound processing mechanisms than is usually considered as 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. The in-channel cancellation model fits this requirement *qualitatively*. Quantitative predictions require a more complex model and assumptions, which are outside the scope of this paper.

The tone/noise masking asymmetry has previously 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 superimposed on a pure or harmonic tone. 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 using a model on the basis of a modulation filterbank (Dau et al 1991): the addition of a probe with wider bandwidth excites modulation filters of higher modulation frequency. To account for such results, the in-channel cancellation model would need to invoke the *approximate periodicity* of a narrowband noise relative to a wider-band masker. Whether this explanation can explain the results quantitatively remains to be determined. Additional detection cues may also be involved, such as distortion products, beats, or dip listening (Moore et al 1998).

In-channel vs harmonic cancellation

A similar asymmetry is observed for harmonic maskers relative to inharmonic or noise-like maskers (reviewed by de Cheveigné 2021). That asymmetry could be accounted for by the *harmonic cancellation model* (de Cheveigné 1993, de Cheveigné 2021) that is closely related

to in-channel cancellation. In the harmonic cancellation model, all peripheral channels are filtered by a cancellation filter with the same delay parameter T, whereas here T is set independently in each channel. In both models, these parameters are set automatically by a data-driven mechanism, so they do not differ in their number of free parameters.

The in-channel cancellation model can be seen as a natural extension of the harmonic cancellation model, analogous to the modified EC models (Culling and Summerfield 1994; Breebart and Kohlrausch 2001; Ackeroyd 2004) that extend Durlach's (1963) original EC model. This extension may allow the model to account for data that suggest that inharmonic maskers may share properties of harmonic maskers (e.g. Roberts and Brunstrom 2001; Deroche et al 2014; Popham et al 2018). Whereas harmonic cancellation (in its standard formulation) would be confused by the lack of a clear "period" for such maskers, in-channel cancellation can be effective either because the inharmonic spectrum may be approximated by a harmonic series over a limited frequency range, or because that spectrum is sparse, allowing partials to be cancelled individually, or both.

Spectral sparsity was also invoked by Deroche et al (2014) as a condition for spectral glimpsing: a target partial can be glimpsed between *adjacent* masker partials if their spacing is sufficient. In-channel cancellation of partials also benefits from sparsity, but it has a slightly different requirement: the spacing between *next-to-adjacent* partials must be sufficient, so that a masking partial can be cancelled within channels that it dominates (as for the low-rank partials in Fig. 2 middle).

Neural implementation

Two questions must be addressed: how might the linear filtering operations required by the model be implemented`? Can such processing 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 reviews 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.

One might wonder whether the cascade of filters required by the model can be effectively approximated given (a) non-linear transduction, (b) the stochastic representation of signals as the instantaneous probability of spike firing. To settle this issue would require simulation with realistic models of transduction and neural processing, which is beyond the scope of this paper. However, there is some reason for optimism. The cancellation filter exploits *periodicity*, which is preserved by non-linear transduction, and it has been shown in previous simulations (de Cheveigné 1993, Guest and Oxenham 2019) that a stochastic spike-based neural implementation can offer a functional approximation of cancellation (as also argued for the binaural EC model, e.g. Franken et al 2021). Nevertheless, a more complete simulation is required to provide quantitative predictions.

Predictive coding, invariance, unconscious inference

In-channel cancellation implements a form of *predictive coding* (Barlow and Rosenblith 1961; Friston 2018) 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, and an "error signal" that supports unmasking of a weaker concurrent target. In-channel cancellation is related to linear predictive coding (LPC) (Atal 2006), and one might speculate that more complete forms of LPC could be approximated with some functional benefit (pursuing this idea is beyond the scope of this paper).

One can argue that a major goal of auditory processing is to ensure *invariance*, both to irrelevant stimulus dimensions and features for the purpose of classification, and to competing sound sources for the purpose of auditory scene analysis (de Cheveigné 2021). Invariance to interfering sounds is necessary so that a relevant sound source can be attended to as effectively as if the interference were not present. One can interpret well-known auditory processing abilities such as cochlear frequency resolution or temporal resolution as serving this goal, 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 wideband noise (Fig. 2 top), and in-channel cancellation extends this ability to a situation (tonal masker of an on-frequency probe) for which cochlear filtering does not suffice (Fig. 2 bottom).

Cochlear filtering ensures invariance with respect to an interfering sound by allowing channels dominated by interference to be ignored. However, parts of the target that fall in those channels are then missing: the target is incomplete. The same is true for cancellation: parts of the spectrum that fall on zeros of the filter (Fig. 3, right) are missing. Missing data can be accommodated via a process of unconscience inference (Helmholtz 1867) by which internal models are constrained by the available (albeit incomplete) information. See de Cheveigné (2021) for a more complete argument.

Pitch, pitch change, transparency, tonality

A by-product of cancellation is the automatically-determined delay parameter. For the harmonic cancellation model, it served as an estimate of the period of a harmonic stimulus, and thus a cue to its pitch (de Cheveigné 1998). The in-channel cancellation model differs in that it offers *multiple* delay estimates, each corresponding to the period of a partial or groups of harmonically-related partials that dominate each 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 (similar to the "spectral pitches" invoked by Terhardt 1974). They are of interest if they concord (as for a harmonic complex), but less so if they do not (as or an inharmonic complex), at least for tasks that require judging *the* pitch of a stimulus.

However, they might still support tasks that involve *pitch change*. Demany and Ramos (2005, Demany et al 2011) found that subjects could detect the direction of pitch change for one partial within an a 5-partial complex, despite the fact that partial's pitch was not salient before the change. McPherson et al (2018) found that subjects could judge pitch change between inharmonic complex sounds as well as between harmonic complexes (at least for short delays between sounds). This suggests the existence of "frequency shift detectors" as reviewed by Demany and Semal (2018).

The in-channel cancellation model can be invoked to support an array of in-channel frequency shift detectors, for example as schematized in Fig. 14. Focusing on one channel, and assuming an array of cancellation filters indexed by delay within that channel, the direction of shift can be found by observing its output to the left of the original minimum (shorter delays). A decrease indicates an upward frequency shift, an increase indicates a downward shift (other strategies are of course possible). The outcome (upward vs downward shift) can be summarized across channels by an averaging or voting mechanism (e.g. to explain McPherson et al 2018), or by attending to channels for which there is change (e.g. to explain Demany and Ramos 2005).

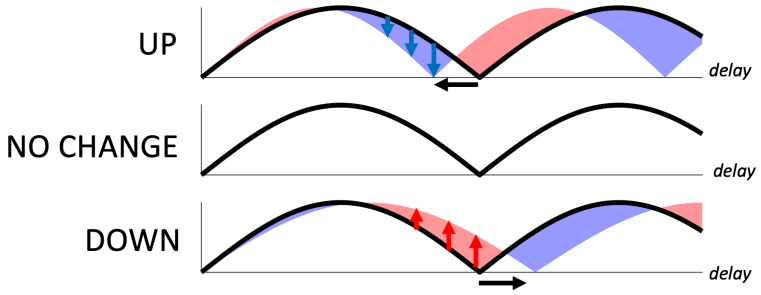

Fig. 14. Schematic principle of an in-channel frequency shift detector. Each plot represents the amplitude at the output of a cancellation filter as a function of its delay parameter (or across an array of such filters indexed by delay) at two moments in time. For an upward change (top), the amplitude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

Importantly, if in-channel cancellation were replaced by global harmonic cancellation, the shift detector would fail for stimuli that do not produce a clear output-vs-delay pattern (e.g. an inharmonic complex). The in-channel cancellation model is an improvement in this respect.

A mechanism that can suppress sounds that are approximately harmonic or spectrally sparse implies for such sounds a property of *transparency* relative to other sounds. This property is relevant for example for perceptual coding of audio signals, as it determines the degree to which quantization noise is masked (Johnston 1988). It may also be useful in a musical context that involves building a complex sound structure with multiple elements (polyphony).

Subjectively, sounds such as inharmonic tones, or bandpass noise, are often judged to have a pitch-like quality (sometimes referred to as "tonality"). In-channel cancellation might offer the basis for metrics to qualify or quantify such a quality.

Spectro-temporal fan-out

In-channel cancellation relies on cochlear filtering to create the channels, and time-domain processing to perform the cancellation. In this it 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), and it has been suggested that the delays required by time-domain processing (for example to implement a long impulse response) might themselves result from between-channel interactions (de Cheveigné and Pressnitzer 2006). A combination of cochlear filtering and time-domain processing is integral to all these models.

According to those models, auditory frequency selectivity is *not* determined entirely by the shape of cochlear filters, since filters of different shapes can be synthesized via subsequent time-domain processing (as exemplified in Fig. 3). This intuition can be formalized (assuming linearity) by modelling the cochlear filter bank as a M-column matrix $\boldsymbol{M}$ of finite impulse responses of order N, applicable to the N-column matrix $\boldsymbol{X}$ of delayed stimulus signals (delays $1 \cdots N$). The matrix $\boldsymbol{Y}$ of cochlear-filtered signals (one channel per column) can then be interpreted as the product of $\boldsymbol{X}$ by $\boldsymbol{M}$. If the matrix $\boldsymbol{M}$ allows an *inverse*, (this requires that its rank be larger than N), multiplying $\boldsymbol{Y}$ by this inverse would reconstruct $\boldsymbol{X}$, which could then be multiplied by an arbitrary impulse response $\boldsymbol{h}$ of order N to obtain an arbitrarily-filtered version of the stimulus. In other words, *any filter* (of order at most N) applicable to the acoustic stimulus can be implemented instead 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.

From this reasoning, we might be tempted to conclude that the limits of cochlear filter selectivity are irrelevant to auditory frequency resolution. That conclusion should be tempered, given the non-linearity and limited dynamic range of transduction and neural

processing. From measurements of binaural or harmonicity-based unmasking (reviewed in de Cheveigné 2021, p10) one can expect a benefit of 3 to 15 dB from post-cochlear filtering. Cochlear filtering is required for greater attenuations, and thus the concepts of "critical band" and "resolvability" remain entirely relevant.

We assumed that the auditory brain has access to both the cancellation-filtered and unfiltered signal within each peripheral channel. This implies a fanout by a factor 2, and mechanisms such as LIN, phase-opponency, binaural interaction, etc., performed concurrently, would lead to a larger factor. Non-linear transforms applied to these, such as demodulation, onset or offset detection and so on would augment the fanout by an additional factor. This is consistent with the massive fanout observed in the auditory brainstem and subsequent relays, by which the ~30000 auditory nerve fibers feed into millions of neurons at the level of auditory cortex.

Functionally, such a fanout is useful for classification, as a complex decision boundary within a feature space can be implemented by applying a linear classifier to the higher-dimensional space of transforms of those features (Duda et al 2012). Random transforms are sufficient, 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 or designing a smaller set of "a priori useful" transforms. It was argued earlier that an important goal is to attain *invariance* to the presence of interfering sound sources. 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" of transforms within the fanout. In-channel cancellation would be a "good thing to have" for the auditory brainstem.

Variants of the cancellation filter

To cancel a partial of frequency f, a delay equal to any multiple $T = k/f$ of its period is equally effective. Indeed, 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.

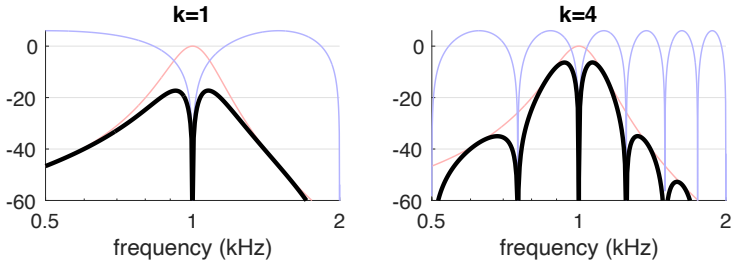

Fig. 15. Transfer functions of compound filter with CF=1 kHz for $T = 1$ ms (left) or $T = 4$ ms (right). Both offer perfect rejection at 1 kHz, but the filter on the right imposes less attenuation on the spectral region close to 1 kHz.

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 that case $k = 1$ might be preferable. Furthermore, if the situation requires a larger delay so as to cancel multiple harmonically-related partials within a channel, it might be worthwhile, to avoid long delays, to approximate the cancellation filter as the 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, essentially all the power is accounted for by just two partials within all channels up to about 1 kHz. A two-filter cascade would be effective to cancel them. These variants (and others) 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 and/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. This estimate can be used as a cue to pitch, or to detect frequency change between inharmonic stimuli that do not evoke an unambiguous 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.

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.

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.

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.

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.

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.