Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Phonetics 31 (2003) 547–561
Journal of Phonetics
www.elsevier.com/locate/phonetics

Time-domain auditory processing of speech

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

Received 18 September 2002; received in revised form 27 May 2003; accepted 10 June 2003

Abstract

This paper reviews evidence for central processing of temporal patterns, describes a number of models that perform useful functions on the basis of temporal processing, and aims to offer insight into mechanisms and constraints, in particular with respect to temporal resolution. The focus is on the region of short duration where temporal descriptions overlap with spectral descriptions. Time-domain cues are used by the listener to parse the auditory scene and assign the relevant fragments to one speaker among several. The cues include interaural delays (that vary according to source position), periodicity (of voiced speech) and envelope modulations such as onsets. Temporal cues may also determine attributes such as pitch (intonation) and timbre (of vowels or consonants), that are often described in terms of spectral cues. Spectrum and time are closely linked. With a frequency-selective cochlea, the ear is equipped to analyze spectra, but from physiology we know that the temporal counterpart of spectral patterns is also present in the auditory nerve. Temporal accuracy decreases as patterns proceed towards the cortex, but precise patterns have been measured at several levels, and there is abundant evidence for neural "circuitry" specialized for time analysis. Several models implement useful functions on the basis of temporal processing. Pitch may be explained on the basis of interval statistics between nerve firings (although consensus on this explanation is not complete), and similar models exist for vowel timbre identification. Segregation of competing voices may be explained by models that use time to either label frequency channels (created in the cochlea) as belonging to one voice or another, or to tease apart information within each channel. An important aspect of these models is that they defer to the central nervous system operations that are often thought to occur at the periphery. If this account is correct, frequency and temporal resolution are not entirely determined by properties of the cochlea.

*Tel.: +33-1-4478-4846; fax: +33-1-4478-1540.
E-mail address: cheveign@ircam.fr (A. de Cheveigné).

0095-4470/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0095-4470(03)00041-X

1. Introduction

Time is essential to speech. It makes sense to sample a visual scene at a certain instant, but not a sound scene: time must flow for sound to exist. Patterns relevant to speech extend over a very wide range of time scales. Interaural time differences of about 10 μs–1 ms are used to spatially organize the sound scene and attend to a speaker. Periodicities in the 400 μs–5 ms range are cues to vowel identity (Peterson & Barney, 1952), a range that overlaps with the 250 μs–30 ms range of musical pitch that serves also for intonation (Semal & Demany, 1990; Pressnitzer, Patterson, & Krumbholz, 2001). Further along this scale we might find cues to segmental identity and prosodic structure (Rosen, 1992), and yet further are memory and context effects, development and learning, the evolution of the language, and the evolution of the species. Each of these phenomena involves an accumulation (or integration) of events or patterns over time. Units of integration are large at one end of the scale (lifetime of a species, language, or individual), and small at the other (duration of a phrase, segment or glottal cycle). The time scale that characterizes a pattern (for example the period of a periodic sound), and the time scale over which information about that pattern is accumulated, need not be equal. They are, however, likely to be commensurate.

Features in the timbre or pitch range are often described in terms of *spectrum*. There are three basic reasons for that, two good and one bad. The first (good) reason is that the cochlea approximates a Fourier transform, different regions of the cochlea being responsive to different spectral regions within a sound. The second (good) reason is that the Fourier transform decomposes signals into sinusoidal waves, which have a special status with respect to systems that are linear and time-invariant: a sinusoidal wave at the input produces a sinusoidal wave at the output. Its *amplitude* and *phase* are changed, but the frequency stays the same, and the shape is still sinusoidal. Thus, decomposing a sound into its various frequency components, feeding each through the linear system, and adding up the outputs, is a convenient way of predicting how any sound is affected when it goes through the system. Many systems that produce, transmit or process sound are linear and time-invariant.

The third (bad) reason stems from a misinterpretation of the well-known theorem that says that a periodic signal of period T_0 can be decomposed into a sum of components with frequencies at integer multiples of a common "fundamental" frequency $F_0 = 1/T_0$. The theorem does *not* say that a sinusoidal component of frequency F_0 should be expected. The assumption that pitch is determined by this fundamental component has led to much effort being invested in the "missing fundamental" problem, which in my view is fundamentally a missing problem. The property relevant for pitch is periodicity, which is easy to characterize if the theorem is *not* used. There is nothing in Fourier's theorem to imply that periodicity is best characterized by the shape of the spectrum. It can be done (Fourier transform of a long enough chunk of waveform, followed by matching of the spectrum to some model of a harmonic spectrum), but it is not the easiest way of doing things.

Setting aside this third reason, the first two fully justify the use of frequency descriptions for pitch and timbre. With respect to the first we should remember that the cochlea is not a perfect Fourier analyzer. With respect to the second, we must remember that systems involved in speech production and perception are not in every respect linear and time-invariant.

If the cochlea performed a perfect short-term Fourier analysis, the patterns sent to the brain would be slowly varying spectra, and temporal patterns (on the time scale relevant for pitch or

timbre) would stop at the cochlea. Actually the cochlea is more like a bank of filters with unsmoothed outputs, and thus it does not prevent temporal patterns from leaking through to the brain. The next section reviews evidence for temporal patterns within the nervous system, not only within the auditory nerve just after the cochlea, but also at multiple relays within the brainstem, where there is evidence for neural circuitry specialized for the transmission and processing of temporally accurate patterns.

2. Specializations for time in the auditory system

The cochlea responds best to high frequencies at its base and low frequencies at its apex, and this orderly distribution is projected centrally (du Verney, 1683). Tonotopic distributions of neural activity are found at all levels from cochlea to cortex, and theories of auditory processing are often based on such slowly varying spectral representations.

However, recordings from single fibers of the auditory nerve of animals show a clear temporal structure in response to laboratory, environmental, or speech sounds (Kiang, 1965; Delgutte & Kiang, 1984a, b). Responses within each fiber consist of "spikes" that seem to occur at random, but with a probability that accurately follows the temporal patterns of sound as modified by cochlear filtering and hair cell transduction. The highest frequency at which temporal structure is measurable is about 3–5 kHz in the cat. It is lower in the guinea pig (Russel & Palmer, 1986), higher in the Barn Owl (9 kHz, Köppl, 1997), and unknown in humans. It has been argued that temporal structure can remain useful at frequencies beyond the cutoff, for example up to 10 kHz assuming data from the cat (Heinz, Colburn, & Carney, 2001).

In addition to the auditory nerve, accurate temporal responses are found at levels up to the inferior colliculus: cochlear nucleus, olivary complex, and nuclei of the lateral lemniscus. The upper limit of synchrony to sustained tones or amplitude modulation gets progressively lower as one proceeds. It is typically 3 kHz in the cochlear nucleus, 1.2 kHz in the inferior colliculus and 100 Hz in the cortex (Ehret & Romand, 1997). These figures do not necessarily imply an inverse trend of temporal resolution. For example, latencies of cortical responses to irregularities in the stimulus show a scatter as little as 1 ms (Hiel & Irvine, 1997). It is likely that the drop in synchrony is the result of a limited capacity to entrain to repetitive stimuli, rather than poor temporal accuracy.

There is evidence for neural "hardware" (or "wetware") specialized for the transmission of temporal patterns (Oertel, 1999). Myelinated axons conduct impulses faster, and possibly with less jitter, than nonmyelinated axons. Axons of auditory nerve fibers, projections of relay neurones of the cochlear nucleus ("bushy" and "octopus" cells), and inhibitory projections of the medial nucleus of the trapezoidal body (MNTB) are myelinated. Calyce-type synapses ensure transmission of spikes from the presynaptic to the postsynaptic neurone with high reliability and low jitter. Such synapses are found between auditory-nerve axons and bushy cells in the cochlear nucleus, between projections of these bushy cells and MNTB neurones, and between projections of octopus cells and neurones of the nuclei of the lateral lemniscus (Schwartz, 1992; Joris, 1996). This last pathway seems to be particularly important in man (Adams, 1997). Finally, there is evidence that cell membranes and neurotransmitter receptors are specialized to speed up neural transmission, reduce jitter, and shorten recovery times (Sabatini & Regehr, 1999; Oertel,

1999). Given their cost in terms of metabolism and evolutionary tradeoffs, these specializations would probably not exist without a functional role useful for survival.

3. Processing principles

Given the evidence for temporal patterns in the auditory nervous system, we need detailed models to explain how they are processed, and what is gained by such processing. Many ideas have been put forward, but before reviewing them it is worth saying a few words about the patterns themselves.

3.1. Temporal patterns

Temporal patterns can take various forms. The regular repetition of the same event (for example glottal closure) constitutes a periodic pattern indicative of voicing, while a succession of events of different nature (for example voice onset and burst of a plosive) may signal the occurrence of a particular phoneme. In both cases, the arrangement of events in time constitutes the pattern. We must thus assume that each event is located in time, so that one can speak of temporal regularity or temporal order. However an event itself has a temporal extent, or *support*, over which the pattern characteristic of the event develops. This can be understood as the segment of time over which an event-recognizing mechanism would gather information to detect the event (and distinguish it from other events). Voice onset occurs at an instant, but, to detect this instant, one must take into account the silence that precedes it and the regular pattern of vibration that follows it. A wider context may be relevant, for example to allow normalization by speech rate or ambient noise, and in this sense the support of an event is potentially quite large.

The regular repetition of an event constitutes a periodic pattern. Thus, one might wish to characterize periodicity in terms of events (such as zero-crossings, or peaks, or glottal pulses). However, there are at least two alternatives. One is to use a measure of "self-similarity" such as the autocorrelation function (ACF). The ACF works by comparing the waveform to itself for various delays, or "lags". For a periodic waveform, similarity is greatest for a lag equal to the period, and this leads to a peak in the ACF that may be used as a cue to periodicity (Section 3.2). A second alternative is to use a set of regular patterns as "yardsticks". Sines and cosines can be used for that purpose: each is compared to the waveform by taking their vector product (i.e. multiplying term by term and adding), an operation known as the Fourier transform. Neither of these alternatives involves events.

What is the minimum duration needed to detect that a pattern is periodic with period T? Obviously it cannot be less than T, but is T enough? If periodicity is defined in terms of events, one needs at least two events, and to recognize that they are indeed repetitions of the *same* event, the pattern must include their *supports*. If periodicity is defined using autocorrelation, then one needs to include, in addition to a lag of at least T, the *integration* time of the ACF (see Eq. (1) in Section 3.2). If periodicity is defined in terms of the Fourier transform, one needs an interval large enough so that the spectral resolution of the transform tells us whether the spectrum is harmonic or not. In each case, the minimum segment is greater than T (a more precise estimate is given in Section 4.1).

The point made here is that the interval of time required to characterize a feature of the speech waveform is at least as large as the temporal dimensions of the feature, and often larger. Intervals of neighboring events may overlap. In some ways the situation is analogous to that encountered with more abstract "events" such as phonemes or words, that also involve integration over wide and overlapping segments of time (Plomp, 2002).

3.2. Processing schemes

This section offers an overview of central processing schemes. Each tries to do something useful with the temporal patterns carried by the auditory nerve. Roughly speaking, the pattern carried by each auditory nerve fiber reflects the temporal structure of the speech waveform as modified by the following steps: (a) narrow-band filtering (b), half-wave rectification, (c) mild low-pass filtering representing loss of synchrony at high frequencies, and (d) production of nerve spikes with a probability that follows the speech waveform as modified by the previous steps. Step (b) also includes a form of amplitude normalization (automatic gain control) that reduces the dynamic range of internal patterns relative to that of the stimulus.

A first idea is that each cochlear filter output is followed by a "neural" band-pass filter of the same center frequency, on the assumption that cochlear selectivity is insufficient and that neural sharpening would do some good. This corresponds to the "average localized synchrony rate" (ALSR) of Young and Sachs (1979), or "average localized synchrony measure" (ALSM) of Delgutte (1984), the matched filters of Srulovicz and Goldstein (1983), or the "lateral inhibitory network" (LIN) of Shamma (1985). The peripheral filterbank would thus be shadowed by a central filterbank, to produce tonotopic patterns sharper than those measured in the auditory nerve. Such sharpened patterns have not yet been found in the auditory system.

A second idea is that the output of each channel independently undergoes a Fourier transform, with a frequency axis distinct from the tonotopic axis projected from the cochlea. This corresponds to the dominant component scheme of Delgutte (1984), and to the recently popular idea of modulation spectrum (e.g. Meyer & Berthommier, 1996; Dau, Püschel, & Kohlrausch, 1996; Dau, Kollmeier, & Kohlrausch, 1997). Instead of a sharpened tonotopic pattern, these schemes produce a two-dimensional pattern defined over characteristic frequency (CF) and best modulation frequency (BMF). One might ask what is the functional advantage of a central spectral analyzer, given that a spectral analyzer already exists in the cochlea. Note, however, that they do not apply to the same patterns: cochlear analysis applies to the waveform, whereas central Fourier analysis would apply to the (roughly speaking) half-wave rectified band-pass filtered speech waveform.

A third idea is to calculate correlation functions within each channel. Licklider (1951, 1959) proposed a pitch perception model based on the running ACF. The ACF is obtained by multiplying together samples of the waveform that differ by a certain delay τ, and summing the products over a window:

$$r_t(\tau) = \int_t^{t+W} s(\theta)s(\theta-\tau)\,\mathrm{d}\theta, \tag{1}$$

where t is the instant at which the calculation is performed, s is the waveform, W is the size of the integration window, and θ is an integration variable. The purpose of the window is to limit the amount of waveform that is processed. The result $r_t(\tau)$ is the ACF at lag τ, calculated at time t. Its

value is large if $s(\theta)$ and $s(\theta-\tau)$ are similar, which is the case for a periodic waveform if τ equals the period. Thus, a peak in $r_t(\tau)$ is a cue to the period. Licklider suggested that a similar operation could be applied to neural patterns within the auditory system, using synaptic or conduction delays to implement lag, coincidence-counting neurones to implement multiplication, and postsynaptic temporal summation to implement integration. Jeffress (1948) had earlier proposed interaural crosscorrelation to estimate time-of-arrival differences, for sound source localization. To calculate the crosscorrelation function, one simply modifies Eq. (1) so that it multiplies waveforms s_R and s_L from the two ears:

$$c_t(\tau) = \int_t^{t+W} s_R(\theta) s_L(\theta-\tau)\,\mathrm{d}\theta. \tag{2}$$

The crosscorrelation function shows a peak at an internal interaural delay τ that compensates for the external interaural delay characteristic of the azimuth of the source. This allows the source to be localized.

A fourth idea is to replace the *multiplication* operation involved in autocorrelation (Eq. (1)) or crosscorrelation (Eq. (2)), by *subtraction*. A physiological implementation can be obtained by replacing excitatory by inhibitory neural interaction. This idea was used in the binaural equalization-cancellation (EC) model of Durlach (1963) (Durlach & Colburn, 1978) to explain binaural masking release, and in the monaural harmonic cancellation model of de Cheveigné (1993) to explain segregation on the basis of harmonic structure. Cancellation and autocorrelation are closely related (de Cheveigné, 1998).

Other ideas have been proposed, and more are no doubt lurking in ingenious minds. Patterson, Allerhand, and Guiguère (1995) proposed a "strobed temporal integration" model that is equivalent to extracting events (e.g. one per period) from neural patterns, and cross-correlating the event train with the ongoing pattern (rather than applying autocorrelation to the latter). Cariani (2001) describes recurrent networks within which patterns circulate and are correlated (or convolved) with incoming patterns. Interestingly, temporal codes and processing mechanisms have been proposed for other modalities such as vision (Thorpe, Fize, & Marlot, 1996). Maass (1998) has shown mathematically that neural networks that use time of arrival of spikes are more powerful than those that use average spike rate, "power" being measured by the number of elements required to implement a function. Among other interesting possibilities offered by "spiking neural networks" are filters with arbitrary transfer functions. These may be implemented by adding up appropriately delayed excitatory and inhibitory postsynaptic potentials (EPSPs and IPSPs). While these ideas have mostly been explored in modalities other than hearing, they certainly make sense within this very temporal modality. The next section describes some models that utilize these ideas for specific tasks.

4. Models that use time

4.1. The autocorrelation model of pitch

The idea that pitch might depend on the spacing between pulses within the auditory nerve dates back to the "telephone" and "volley" theories of Rutherford (1886) and Wever and Bray (1930).

Licklider (1959) gave it a more specific formulation on the basis of the interval statistics of neural discharges, similar to the ACF (Section 3.2).

The autocorrelation model accounts for a wide range of pitch phenomena (Meddis & Hewitt, 1991; Cariani & Delgutte, 1996), and it is currently quite popular. Its major strength, with respect to the previous generation of pitch models that derived fundamental periodicity from frequencies (or periods) of individual partials (e.g. Goldstein, 1973), is that it does not require a separate pattern-matching stage. Fundamental periodicity is derived from "resolved" and "unresolved" components according to the same basic mechanism. This feature is also a weakness, as it leads one to expect that pitches of stimuli with resolved and unresolved components are equally salient, which is not the case (e.g. Carlyon & Shackleton, 1994). This is currently a major issue in hearing (Meddis & O'Mard, 1997; Carlyon, 1998). From an applied point of view, autocorrelation has proved to be an effective basis for fundamental frequency estimation of speech and music (de Cheveigné & Kawahara, 2002).

It is interesting to know the shortest duration of a signal necessary to estimate its period. It is the sum of two terms. The first is the range of values of τ that is explored, the second the size of the integration window W (Eq. (1)). A priori, both are determined by the range of *expected* periods rather than the actual period, and both should be at least as large as the largest expected period, T_{MAX}. This corresponds to the familiar rule of thumb: period estimation requires a chunk of signal with duration at least $2T_{MAX}$. It is possible to reduce this to $T + T_{MAX}$, where T is the period being measured (de Cheveigné & Kawahara, 2002), but no shorter. Shorter than T_{MAX} integration implies fluctuating behavior, and a pitch model that makes that assumption cannot guarantee a stable percept of pitch.

Actually, Licklider (1959) or Meddis and Hewitt (1991) assumed integration over a window shaped as an *exponential* (decreasing towards the past), rather than as a square (as in Eq. (1)). This difference is of minor importance: the time constant of the exponential maps roughly to the width W of the square window. They hypothesized a time constant of 2.5 ms, and Meddis and Hewitt (1992) later proposed 10 ms. The smaller value is adequate above 400 Hz but no lower, while the larger is adequate down to 100 Hz. Recent efforts to determine experimentally the size used by the auditory system were reviewed by Wiegrebe (2001). It turns out that the integration window size may be *task dependent*. In particular, it may vary with the F_0 of the stimuli used in a task. Data are consistent with an integration duration of twice the period with a minimum duration of 2.5 ms. The only problem with this proposition is that period-dependent integration assumes prior knowledge of the period, a circular process.

These estimates determine the minimum duration for which tasks involving pitch are possible. However, performance benefits from longer intervals of signal, if available (Moore, 1973). During a stimulus, the summation process can be "reset" by transient events (Plack & White, 2000a), as has also been observed in the binaural system (Hafter & Buell, 1990). It is as if the auditory system can, within limits, tailor available evidence by applying the window that offers best performance (Dau et al., 1996; Moore, 2003). This idea is closely related to a technique recently proposed in automatic speech recognition, called "missing feature theory", in which different parts of spectro-temporal patterns are weighted differently according to their reliability (Cooke, Morris, & Green, 1997).

Integrating over longer windows can reduce noise, leading to better resolution. As another way to improve resolution, it has been proposed that higher-order peaks of the ACF (at 2T, 3T, etc.) are used in addition to the peak at the period (Yost, 1999; Plack & White, 2000b; de Cheveigné, 2000). A given duration can thus be used to maximize either lag range or integration, so the two

factors are confounded in experiments that attempt to probe them. The *maximum* duration of integration used by the auditory system has been estimated at about 210 ms (Grose, Hall, & Buss, 2002). If frequency is changing (as for example in an intonation pattern), shorter windows may be needed to resolve the temporal pattern of pitch, and to limit the degree of nonstationarity within each integration window.

To summarize, the autocorrelation model accounts well for a wide range of pitch phenomena. Its major weakness is that it does not predict the differences in performance observed for stimuli with resolved vs. unresolved components. Processing requires a minimum duration that depends upon the stimulus period but, according to the task, information may be gathered over a longer duration. Functionally, longer windows favor accuracy of the period estimation while shorter windows allow fast modulations to be followed (or at least to be recognized as pitch-like).

4.2. The autocorrelation model applied to timbre

The ACF has also been used to account for the perception of timbre. The ACF is equal to the Fourier transform of the power spectrum, so it is equivalent in terms of information to a spectrum (without phase). Meddis and Hewitt (1992) used template matching of the short-lag portion of the ACF (below 2.5–4 ms) in a model of vowel identification. This is an alternative to the usual assumption that vowel spectra are represented as tonotopic patterns of activity. The motivation for a time-domain model is weaker than in the case of pitch, as peripheral selectivity is sufficiently fine to resolve formant patterns.

A possible functional advantage of an autocorrelation-based mechanism is that different features of the power spectrum are represented (via the Fourier transform) by different parts of the ACF. The short-lag part represents large-scale features (spectral envelope), while the long-lag portion reflects also the fine harmonic structure. Truncation of the ACF at a lag equal to one-half the period is equivalent to sampling the spectrum precisely at harmonics of F_0 (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999; de Cheveigné & Kawahara, 1999). This reduces, to some degree, the interactions between F_0 and spectral shape. If the auditory system were capable of performing an operation of this sort, it would go some way to solving the old problem of F_0-dependency of estimates of spectral shape (Klatt, 1982).

The ACF being equivalent to a *power* spectrum, high-amplitude spectral features such as the first formant are over-represented with respect to lower-amplitude features. This is one reason why the cepstrum (Fourier transform of the *log* spectrum) is preferred for speech processing. However, recall that the transduction step within the cochlea has amplitude normalization properties (Section 3.2). As a result of normalization, the overall pattern of ACFs should reflect spectral features of the stimulus in a relatively balanced fashion.

To summarize, the ACF model can explain how spectral shape (up to at most 5 kHz) might be extracted on the basis of within-channel temporal patterns instead of (or in addition to) across-channel spectral patterns.

4.3. Segregation based on harmonicity

Speech is often heard against a background of other speakers or noise. Among the cues and mechanisms responsible for speech segregation (Darwin & Carlyon, 1995), *harmonic structure* is

useful when competing voices have different fundamental frequencies (Brokx & Nooteboom, 1982). Among the many models proposed to account for this, those based on a spectral representation do not work well, unless they are given a much better resolution than that of the cochlea (Parsons, 1976; Assmann & Summerfield, 1990; de Cheveigné, 1993). This is a task for which temporal processing seems necessary.

Time-domain models work according to either of two principles. With *channel selection* models, the temporal pattern within each channel is used to assign it to one source or another (Meddis & Hewitt, 1992). With *channel splitting* models, each channel is shared between sources (de Cheveigné, 1993, 1997). Channel selection obviously requires peripheral filtering to obtain the channels among which to select. Channel splitting is less dependent on peripheral filtering (although that filtering may make things easier by improving signal-to-noise ratio within individual channels). Channel selection fails if all channels are dominated by the same interfering source, yet experiments show that F_0-guided segregation still occurs in that situation (de Cheveigné, Kawahara, Tsuzaki, & Aikawa, 1997; de Cheveigné, 1999), so channel selection cannot provide a complete account. A likely proposition is that both principles are at work.

Temporal processing is based on the ACF in Meddis and Hewitt's model, on the modulation spectrum in that of Meyer and Berthommier (1996), or on a ''neural cancellation'' filter (followed by autocorrelation) in the model of de Cheveigné (1993, 1997). Yet other temporal processing schemes have been proposed (Assmann & Summerfield, 1990; Brown, Cooke, & Mousset, 1996; Cooke & Ellis, 2001).

A segregation mechanism requires time. Assmann and Summerfield (1994) found that pairs of concurrent vowels with different F_0's were harder to segregate if shortened from 200 to 50 ms. With very short stimuli, McKeown and Patterson (1996) found that *one* vowel of most pairs was identified from as little as one cycle (at a 100 or 200 Hz fundamental). Identification of the second vowel improved gradually with duration, up to 8 cycles, suggesting that it required the help of a segregation mechanism with a certain degree of ''sluggishness''. However, Culling, Summerfield and Marshall (1994) found similar segregation with F_0's that were static or modulated at a rate of 5 Hz, implying relatively fast tracking of harmonic structure.

The models of Meddis and Hewitt (1992) and de Cheveigné (1993, 1997) both use the F_0 of the dominant vowel to retrieve the weaker vowel. Longer stimuli may help to estimate this F_0, thus explaining McKeown and Patterson's results. As for the *asymmetry* of intelligibility of vowels within a pair, it can be understood from the pattern of mutual masking of their spectral cues, mainly formants one and two (de Cheveigné, 1999).

To summarize, harmonicity-based segregation appears to involve temporal processing. Identification of a weaker voice probably requires estimating the fundamental frequency of the stronger voice, a process that takes time.

4.4. Segregation based on binaural cues

Binaural segregation models also use the principles of *channel selection* and *channel splitting*. As an example of the first, Lyon (1983) calculated cross-correlation functions of signals from both ears, within each cochlear frequency channel. Supposing that different sources have different spectral envelopes, some channels must respond preferentially to one source. Cross-correlation functions in those channels show a peak at the internal delay that matches the difference in time of

arrival from that source to the ears. Channels may thus be grouped as belonging to one source or the other. Patterson, Anderson, and Francis (1996) proposed a similar idea. An example of the second principle is the EC (Equalization–Cancellation) model of Durlach (1963) (Durlach & Colburn, 1978). The EC model adjusts (internally) the relative amplitude and delay of signals from the two ears to make them as similar as possible. The equalized signals are then subtracted, and the remainder is used as a cue to the target.

The EC model has recently been revisited by Culling and Summerfield (1995) and Breebaart, van de Par, and Kohlrausch (2001). In Culling and Summerfield's "modified EC" model, equalization and cancellation are performed independently for each frequency channel. The delay and amplitude parameters that determine cancellation are thus allowed to differ from channel to channel. This allows the model to account for a puzzling phenomenon found by Culling and Summerfield (1995): listeners presented with synthetic four-formant stimuli were unable to group together those formants that came from the same direction. According to the original EC model this task should have been easy.

Binaural processes such as described by the EC model tend to be "sluggish" in comparison to monaural processes (Grantham & Wightman, 1979; Kollmeier & Gilkey, 1990; Culling & Summerfield, 1998) However, the time constants vary considerably between tasks, and also between individuals (Akeroyd & Bernstein, 2001).

The EC model accounts well for the two-ear advantage in detection experiments, for which masking level differences (MLDs) can be as high as 15 dB. Unfortunately, situations that give a large binaural benefit for *detection* of simple stimuli do not always give a similar binaural benefit for *intelligibility* of speech. The benefit of spatial cues is not simply the result of instantaneous unmasking, but appears also to involve the cognitive organization of speech parts belonging to competing voices across time (Darwin & Hukin, 1999). This process is not yet well understood.

To summarize, binaural cues include interaural differences in time, level and spectrum. These may contribute to speech intelligibility in two ways: by improving signal to noise ratio either externally (head shadow and pinna directivity) or internally (interaural interaction), and by providing cues of a more cognitive nature that indicate to the listener the structure of the spatial scene. Such effects are an active field of study.

5. Concluding remarks

A first purpose of this paper was to remind us that the spectral analysis of speech sounds, often assumed to be complete at the cochlea, may continue within the nervous system on the basis of temporal patterns carried by the auditory nerve. The spectral and temporal resolution of this analysis is not entirely determined by that of the cochlea. This conclusion seems to suggest that *spectro-temporal excitation patterns* (Moore, 2003) are not a complete description of the information available to the auditory system. Nevertheless, those patterns have been shown to have good predictive power. It may be that temporal processing has properties that map well to a spectral description in terms of cochlear excitation patterns. Time-domain models are less well developed, less authoritative, and less convenient as descriptive tools. Synchrony-based "auditory images" (e.g. Patterson et al., 1995, 1996) are rich representations, but partly for that reason they are not as convenient as spectro-temporal patterns. A wise course may be to stick with the

spectro-temporal patterns as a descriptive tool, while remaining alert for phenomena that escape them, and that may have their basis in more central processing of temporal patterns.

A second purpose was to review a number of temporal models for pitch, timbre or segregation. Some are in competition with models based on cochlear spectrum analysis, others appear to provide the only explanation of the phenomena they address. The review is not exhaustive: many more models exist. The aim was to give insight into the sort of functions that can be supplied, and the palette of mechanisms that might supply them.

A more authoritative account of what actually goes on in the auditory system is way into the future. We know that temporal processing occurs, but it is hard to make sense of the detailed patterns measured by the physiologists, to assign them a functional role, or to relate them precisely to the patterns postulated by the models. This is true for simple phenomena such as pitch or timbre, and all the more so for speech. Progress may come from (a) investigation techniques such as brain imaging or multielectrode recording, (b) progress in behavioral techniques, (c) computer and theoretical models to make sense of the data and guide future investigation, and (d) feedback from engineering applications such as automatic speech recognition. Theories are known to sometimes have a blinding effect, and such may be the case for the theory of an exclusively spectral processing in the cochlea. If so, exploring the alternative hypothesis of temporal processing may benefit our understanding.

A third purpose was to give some indication of time constants, on the basis of functional and/or behavioral considerations. Time is required to measure the scale of a pattern, differentiate it from other (possibly longer) patterns, stabilize the estimates over time, and counter the effects of internal or external noise. These requirements are hard to tease apart experimentally. Functional constraints set lower limits, but the system appears to take advantage of a longer duration when available, and to adaptively tailor information to maximize performance.

Temporal patterns relevant for speech occur over a wide range of scales. This review concentrated on the small-scale end of this range, where temporal descriptions compete with spectral descriptions. The review of Moore (2003) covers auditory phenomena relevant for speech perception over a wider range of scales, including forward and backward masking, temporal order, etc. Time is an essential dimension of each speech pattern, but it is ''overloaded'' with other important roles. Time separates one pattern from the next. Time is irreversible, so a pattern cannot be revisited without some form of memory. Time separates perception from reaction, and causality determines their order. All these aspects must be taken into account in the design of models to process time.

Acknowledgements

Thanks to Brian Moore, an anonymous reviewer, and the Editors for useful comments on a previous draft. This work was supported in part by the Cognitique programme of the French Ministry of Research.

References

Adams, J. C. (1997). Projections from octopus cells of the posteroventral cochlear nucleus to the ventral nucleus of the lateral lemniscus in cat and human. *Auditory Neuroscience*, *3*, 335–350.

Akeroyd, M. A., & Bernstein, L. R. (2001). The variation across time of sensitivity to interaural disparities: Behavioral measurements and quantitative analyses. *Journal of the Acoustical Society of America*, *110*, 2516–2526.

Assmann, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, *88*, 680–697.

Assmann, P. F., & Summerfield, Q. (1994). The contribution of waveform interactions to the perception of concurrent vowels. *Journal of the Acoustical Society of America*, *95*, 471–484.

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

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, *10*, 23–36.

Brown, G. J., Cooke, M., & Mousset, E. (1996). Are neural oscillations the substrate of auditory grouping? In: Ainsworth, W., & Greenberg, S. (Eds.), *Proceedings of the ESCA Workshop on the auditory basis of speech perception*, Keele (pp. 174–179).

Cariani, P. (2001). Neural timing nets for auditory computation. In S. Greenberg, & M. Slaney (Eds.), *Computational models of auditory function* (pp. 233–247). Amsterdam: IOS Press.

Cariani, P. A., & Delgutte, B. (1996). Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *Journal of Neurophysiology*, *76*, 1698–1716.

Carlyon, R. P. (1998). Comments on "A unitary model of pitch perception" [J. Acoust. Soc. Am. 102, 1811–1820 (1997)]. *Journal of the Acoustical Society of America*, *104*, 1118–1121.

Carlyon, R. P., & Shackleton, T. M. (1994). Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms? *Journal of the Acoustical Society of America*, *95*, 3541–3554.

Cooke, M., & Ellis, D. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, *35*, 141–177.

Cooke, M., Morris, A., & Green, P. (1997). Missing data techniques for robust speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech & Signal Processing* (pp. 863–866).

Culling, J. F., & Summerfield, Q. (1995). Perceptual segregation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *Journal of the Acoustical Society of America*, *98*, 785–797.

Culling, J. F., & Summerfield, Q. (1998). Measurement of the binaural temporal window using a detection task. *Journal of the Acoustical Society of America*, *103*, 3540–3553.

Culling, J. F., Summerfield, Q., & Marshall, D. H. (1994). Effects of simulated reverberation on the use of binaural cues and fundamental frequency differences for separating concurrent vowels. *Speech Communication*, *14*, 71–95.

Darwin, C. J., & Carlyon, R. P. (1995). Auditory grouping. In B. C. J. Moore (Ed.), *Handbook of perception and cognition: Hearing* (pp. 387–424). New York: Academic Press.

Darwin, C. J., & Hukin, R. W. (1999). Auditory objects of attention: The role of interaural time differences. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 617–629.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *Journal of the Acoustical Society of America*, *102*, 2906–2919.

Dau, T., Püschel, D., & Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *Journal of the Acoustical Society of America*, *99*, 3615–3622.

de Cheveigné, A. (1993). Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America*, *93*, 3271–3290.

de Cheveigné, A. (1997). Concurrent vowel identification III: A neural model of harmonic interference cancellation. *Journal of the Acoustical Society of America*, *101*, 2857–2865.

de Cheveigné, A. (1998). Cancellation model of pitch perception. *Journal of the Acoustical Society of America*, *103*, 1261–1271.

de Cheveigné, A. (1999). Vowel-specific effects in concurrent vowel identification. *Journal of the Acoustical Society of America*, *106*, 327–340.

de Cheveigné, A. (2000). A model of the perceptual asymmetry between peaks and troughs of frequency modulation. *Journal of the Acoustical Society of America*, *107*, 2645–2656.

de Cheveigné, A., & Kawahara, H. (1999). Missing data model of vowel perception. *Journal of the Acoustical Society of America*, *105*, 3497–3508.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, *111*, 1917–1930.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., & Aikawa, K. (1997). Concurrent vowel identification I: Effects of relative level and F0 difference. *Journal of the Acoustical Society of America*, *101*, 2839–2847.

Delgutte, B. (1984). Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds. *Journal of the Acoustical Society of America*, *75*, 879–886.

Delgutte, B., & Kiang, N. Y.-S. (1984a). Speech coding in the auditory nerve: I. Vowel-like sounds. *Journal of the Acoustical Society of America*, *75*, 866–878.

Delgutte, B., & Kiang, N. Y.-S. (1984b). Speech coding in the auditory nerve: V. Vowels in background noise. *Journal of the Acoustical Society of America*, *75*, 908–918.

Durlach, N. I. (1963). Equalization and cancellation theory of binaural masking-level differences. *Journal of the Acoustical Society of America*, *35*, 1206–1218.

Durlach, N. I., & Colburn, H. S. (1978). Binaural phenomena. In E. C. Carterette, & M. P. Friedman (Eds.), *Handbook of perception*, Vol. IV. (pp. 365–466). New York: Academic Press.

Ehret, G., & Romand, R. (1997). *The central auditory system*. New York: Oxford University Press.

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, *54*, 1496–1516.

Grantham, W., & Wightman, F. L. (1979). Detectability of a pulsed tone in the presence of a masker with time-varying interaural correlation. *Journal of the Acoustical Society of America*, *65*, 1509–1517.

Grose, J. H., Hall III, J. W., & Buss, E. (2002). Virtual pitch integration for asynchronous harmonics. *Journal of the Acoustical Society of America*, *112*, 2956–2961.

Hafter, E. R., & Buell, T. N. (1990). Restarting the adapted binaural system. *Journal of the Acoustical Society of America*, *88*, 806–812.

Heinz, M. G., Colburn, H. S., & Carney, L. H. (2001). Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve. *Neural Computation*, *13*, 2273–2316.

Hiel, P., & Irvine, D. R. F. (1997). First-spike timing of auditory-nerve fibers and comparison with auditory cortex. *Journal of Neurophysiology*, *78*, 2438–2454.

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative Physiology and Psychology*, *41*, 35–39.

Joris, P. X. (1996). Envelope coding in the Lateral Superior Olive. II. Characteristic delays and comparison with responses in the Medial Superior Olive. *Journal of Neurophysiology*, *76*, 2137–2156.

Köppl, C. (1997). Phase locking to high frequencies in the auditory nerve and cochlear nucleus magnocellularis of the barn owl *Tyto alba*. *Journal of Neuroscience*, *17*, 3312–3321.

Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, *27*, 187–207.

Kiang, N. Y.-S. (1965). *Discharge patterns of single fibers in the cat's auditory nerve*. MIT research monograph, Vol. 35. Cambridge, MA: MIT Press.

Klatt, D. H. (1982). Speech processing strategies based on auditory models. In R. Carlson, & B. Granström (Eds.), *The representation of speech in the peripheral auditory system* (pp. 181–196). Amsterdam: Elsevier.

Kollmeier, B., & Gilkey, R. (1990). Binaural forward and backward masking: Evidence for sluggishness in binaural detection. *Journal of the Acoustical Society of America*, *87*, 1709–1719.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. *Experientia*, *7*, 128–134.

Licklider, J. C. R. (1959). Three auditory theories. In S. Koch (Ed.), *Psychology, a study of a science* (pp. 41–144). New York: McGraw-Hill.

Lyon, R. F. (1983–1988). A computational model of binaural localization and separation. In W. Richards (Ed.), *Natural computation* (pp. 319–327). Cambridge, MA: MIT Press.

Maass, W. (1998). On the role of time and space in neural computation. *Lecture Notes in Computer Science*, *1450*, 72–83.

McKeown, J. D., & Patterson, R. D. (1996). The time course of auditory segregation: Concurrent vowels that vary in duration. *Journal of the Acoustical Society of America*, *98*, 1866–1877.

Meddis, R., & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *Journal of the Acoustical Society of America*, *89*, 2866–2882.

Meddis, R., & Hewitt, M. J. (1992). Modeling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, *91*, 233–245.

Meddis, R., & O'Mard, L. (1997). A unitary model of pitch perception. *Journal of the Acoustical Society of America*, *102*, 1811–1820.

Meyer, G., & Berthommier, F. (1996). Vowel segregation with amplitude modulation maps: a re-evaluation of place and place-time models. In *Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception*, Keele (pp. 212–215).

Moore, B. C. J. (1973). Frequency difference limens for short-duration tones. *Journal of the Acoustical Society of America*, *54*, 610–619.

Moore, B. C. J. (2003). Temporal integration and context effects in hearing. *Journal of Phonetics*, *31*; doi: 10.1016/S0095-4470(03)00011–1.

Oertel, D. (1999). The role of timing in the brain stem auditory nuclei of vertebrates. *Annual Review of Physiology*, *61*, 497–519.

Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, *60*, 911–918.

Patterson, R. D., Allerhand, M., & Guiguère, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *Journal of the Acoustical Society of America*, *98*, 1890–1894.

Patterson, R., Anderson, T. R., & Francis, K. (1996). Binaural auditory images and a noise-resistant, binaural auditory spectrogram for speech recognition. In *Proceedings of the ESCA Workshop on the auditory basis of speech perception*, Keele (pp. 245–252).

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175–184.

Plack, C. J., & White, L. J. (2000a). Perceived continuity and pitch perception. *Journal of the Acoustical Society of America*, *108*, 1162–1169.

Plack, C. J., & White, L. J. (2000b). Pitch matches between unresolved complex tones differing by a single interpulse interval. *Journal of the Acoustical Society of America*, *108*, 696–705.

Plomp, R. (2002). *The intelligent ear*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pressnitzer, D., Patterson, R. D., & Krumbholz, K. (2001). The lower limit of melodic pitch. *Journal of the Acoustical Society of America*, *109*, 2074–2084.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London, Series B*, *336*, 367–373.

Russel, I., & Palmer, A. (1986). Filtering due to the inner hair-cell membrane properties and its relation to the phase-locking limit in cochlear nerve cells. In B. C. J. Moore, & R. D. Patterson (Eds.), *Auditory frequency selectivity* (pp. 199–207). New York: Plenum Press.

Rutherford, W. (1886). A new theory of hearing. *Journal of Anatomical Physiology*, *21*, 166–168.

Sabatini, B. L., & Regehr, W. G. (1999). Timing of synaptic transmission. *Annual Review of Physiology*, *61*, 521–542.

Schwartz, I. R. (1992). The superior olivary complex and lateral lemniscal nuclei. In D. B. Webster, A. N. Popper, & R. R. Fay (Eds.), *The mammalian auditory pathway: Neuroanatomy* (pp. 117–167). New York: Springer.

Semal, C., & Demany, L. (1990). The upper limit of musical pitch. *Music Perception*, *8*, 165–176.

Shamma, S. A. (1985). Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of the Acoustical Society of America*, *78*, 1622–1632.

Srulovicz, P., & Goldstein, J. L. (1983). A central spectrum model: A synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum. *Journal of the Acoustical Society of America*, *73*, 1266–1276.

Thorpe, S., Fize, F., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522.

du Verney, J. G. (1683). *Traité de l'organe de l'ouie, contenant la structure, les usages et les maladies de toutes les parties de l'oreille*. Paris.

Wever, E. G., & Bray, C. W. (1930). The nature of acoustic response: The relation between sound frequency and frequency of impulses in the auditory nerve. *Journal of Experimental Psychology*, *13*, 373–387.

Wiegrebe, L. (2001). Searching for the time constant of neural pitch integration. *Journal of the Acoustical Society of America*, *109*, 1082–1091.

Yost, W. A. (1999). *Pitch-strength discrimination involving regular interval stimuli*. Association for Research in Otolaryngology abstract 232.

Young, E. D., & Sachs, M. B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *Journal of the Acoustical Society of America*, *66*, 1381–1403.