Time-domain auditory processing of speech

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

Abstract

Speech patterns develop over time. Temporal phenomena extend from the very short to the very long: timbre, pitch and segregation cues, segmental, prosodic and rhetoric patterns, memory, learning, development and evolution. This paper focuses on the lower end of the scale, where time-domain descriptions merge with spectral descriptions. Temporal cues are used by the listener to organize the auditory scene, parse acoustic information, and assign the relevant fragments to one speaker among several. Cues include interaural delays (that vary according to source position), periodicity (of voiced speech), and envelope modulations such as onsets. Temporal cues may determine pitch (intonation) and timbre (of vowels or consonants), for which it is also common to find accounts based on spectral cues. Frequency and time are closely linked. With a frequency-selective cochlea the ear is equipped to analyze spectral patterns, but from physiology we know that their temporal counterpart is also present in the auditory nerve. Synchrony to the acoustic signal degrades as patterns proceed towards the cortex, but accurate temporal information is available for processing at several levels and there is abundant evidence for neural "circuitry" specialized for time. Models that explain pitch on the basis of interval statistics between nerve firings are currently popular (although consensus is not complete), and similar models exist for vowel timbre identification. Segregation of competing voices is explained by models that use time cues, either to 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 delay to within the central nervous system processing that is often thought to occur at the periphery. The cochlea is given the role of a bank of "prefilters" rather than that of a Fourier transformer. If this account is correct, frequency and temporal resolution are not entirely determined by properties of the cochlea. The purpose of this paper is to review arguments in favor of central processing of temporal patterns, to describe a number of models that perform useful functions on this basis, and to give some insight into the nature of their mechanisms and their constraints, in particular in terms of temporal resolution.

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 10µs to 1ms are used to spatially organize the sound scene and attend to a speaker. Periodicities in the 400µs to 5ms range are cues to vowel identity (Peterson and Barney, 1952), a range that overlaps with the 250µs to 30ms range of musical pitch that serves also for intonation (Pressnitzer et al., 2001). Further along this scale we might find cues to segmental identity, prosodic structure, rhetorical structure, memory and context effects, development and learning, diachronic phenomena and evolution. Patterns are integrated over large units for the latter (lifetime of a species, language or individual), and small units for the former (milliseconds to hundreds of milliseconds). Integration times are commensurate with the duration of features, but the two are logically distinct.

Features in the timbre or pitch range are often described in terms of *frequency*. 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 regions of the spectrum of a sound. The second (good) reason is that the Fourier transform decomposes signals into sums of 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* change, 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 the 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 the well known theorem that says that a periodic signal can be decomposed into a sum of components at integer multiples of a common "fundamental" frequency. Misinterpretation has led to the concept that a sinusoidal component at that frequency should be expected (the theorem says nothing of the sort), or that periodicity in time is best characterized by the shape of the spectrum (not the easiest way of doing things). Much effort has been wasted on the "missing fundamental" problem, fundamentally a missing problem. Setting aside this third unhappy reason, the first two fully justify the use of frequency descriptions for pitch and timbre. With respect to the first we should remember that cochlear analysis has limited selectivity (not the perfect Fourier analyzer) and does not prevent temporal patterns from leaking through to the auditory nervous system. 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.

1. Specialisations 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 Vernay, 1683). It is often assumed that fast temporal patterns of sound give way in the cochlea to slowly-varying spectral patterns that are processed centrally, as reflected by the tonotopic distributions found at all levels from cochlea to cortex.

However fast temporal patterns do not stop at the cochlea. Recordings from single fibres of the auditory nerve of animals show a clear temporal structure in response to laboratory, environmental or speech sounds (Kiang, 1960; Delgutte et al. 1984a,b). Responses within each fibre 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 haircell transduction. For tones this structure is measurable in the auditory nerve up to about 5 kHz in mammals (9 kHz in the barn owl, K ppl, 1997), which covers much of the range important for speech features. The upper limit of synchrony to tones is progressively lower as one proceeds within the nervous system (3 kHz in the cochlear nucleus, 1.2 kHz in the inferior colliculus, 100 Hz in the cortex), but at all levels up to IC (cochlear nucleus, olivary complex, nuclei of lateral lemniscus) clear temporal responses are observed (Ehret and Romand, 1997).

There is also evidence for neural "hardware" specialized for the transmission of temporal patterns (Oertel, 1999). Myelinated axons conduct impulses faster (and possibly with less jitter) than non-myelinated axons. Axons of auditory nerve fibres, projections of relay neurons 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 presynaptic to postsynaptic neuron 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 neurons, and between projections of octopus cells and neurons of the nuclei of the lateral lemniscus (Joris, 1996; Schwartz, 1992), a pathway that seems to be particularly important in man (Adams, 1997). Finally, there is evidence that characteristics of cell membranes and neurotransmitter receptors are specialized to speed up neural transmission, reduce jitter, and shorten recovery times (Sabatini and Regehr, 1999; Oertel, 1999). Given their cost in terms of metabolism and evolutionary tradeoffs, such specialisations would probably not exist without a functional role useful for survival.

2. Processing principles

We lack models to explain what is gained by this processing, and how. Many ideas have been put forward, but before reviewing them it is worth asking what is meant by "temporal pattern".

2.1 Temporal patterns

A tentative starting point might be to distinguish events, associated with *instants*, and regularities or periodicities associated with *intervals*. Examples of the former are an individual glottal pulse, the burst of a plosive, or the realization of a particular phoneme or word. Examples of the latter are the regularities produced by voicing (fundamental frequency), vocal-tract resonances (formants) or the interaural time delays that characterize the spatial position of a speaker. This distinction is unfortunately not easy to maintain. The definition of a particular event may specify regularity (for example the transition of a formant resonance, or the onset of voicing), while the definition of a regularity may invoke events (for example intervals between glottal pulses).

Events are patterns associated with instants, but the pattern itself develops over time. An event thus has a *support*, the interval over which evidence for the event may be gathered. This raises a number of questions. What is the relation between instant and support? Is the instant itself important, and does it need to be evaluated when an event is recognized and arranged with other events? If not, how might the temporal pattern formed by contemporary events be characterized? What if the supports of different events overlap? Our purpose here is not to give answers to these questions, but to acknowledge that processing schemes must somehow deal with them.

Regularities are associated with intervals, and this seems to bring us back to events, as an interval has a beginning and an end. However there are at least two interpretations of "interval" that don't involve events. One is the ongoing interval or "lag" of the auto- or crosscorrelation function. A pattern is delayed by increasing amounts and compared to itself or to another pattern. If the pattern is regular, the match is better for a certain delay that characterizes the regularity, leading to a peak in the function. As it were, each point of the pattern is given the status of event, intervals between it and every other similar event are measured, and an overall "interval" is derived by averaging or voting among these measurements. A second interpretation of "interval" is the reciprocal of frequency in the Fourier transform. Sines and cosines are used as "yard-sticks" and compared to the pattern by taking a cross-product. A spectrum is simply the degree of match between pattern and yardsticks as a function of their frequency. Defining regularity in terms of events is sometimes convenient, but not indispensable.

Since an event has a support, an event-spotting scheme must *integrate* information over a time interval. Is the interval predefined (by some segmentation process), or should the event-spotter be seen as operating continuously over a sliding window? How large the window, and what should be its shape? What if supports of neighbouring events overlap?

Whatever the answers, it is important to remember that an event-spotting scheme needs to work over a time interval, even if the event is conceptualized as instantaneous. Likewise regularities are spread over time. What is the smallest time window needed to characterize a regularity? If it is defined on the basis of events, it might seem that the smallest window that can contain two events (or even two disjoint windows, one for each event) is sufficient. Thus the period of a periodic pattern might measured over as little as one period of time. We must however remember that each event itself requires a support. Furthermore, determining the characteristic interval (e.g. period) also involves ruling out other intervals, shorter and longer, which means exploring a wider window of time. Finally, the presence of noise (external or internal) may require further temporal integration. The point made here is that the time required to characterize a feature is often greater than the temporal dimensions characteristic of the feature.

2.2 Processing schemes

This section offers an overview of processing ideas for temporal patterns within the auditory system. The next section gives functional models that use some of them.

A first idea is that each cochlear filter output is followed by a "neural" bandpass filter of same center frequency. This corresponds to the "average localized synchrony rate" (ALSR) or "measure" (ALSM) of Young and Sachs (1979) and Delgutte (1984), the matched filters of Goldstein and Srulovicz (1977), 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 to suppose that each channel independently undergoes a Fourier transform, with a frequency axis that does not map to 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. Dau, 1997; Meyer and Berthommier, 1996). A problem with this idea is that it reproduces centrally an operation that is already available peripherally.

A third idea is to calculate correlation functions within each channel. Jeffress (1948) proposed interaural crosscorrelation to estimate time-of arrival differences to localize sound sources. Licklider (1951, 1959) likewise proposed a pitch perception model based on the running autocorrelation function:

$$r_t(\tau) = \int_{-\infty}^{t} w(t-\theta)s(\theta)s(\theta-\tau)d\theta \qquad (1)$$

where $\boldsymbol{s}$ is the signal, $\boldsymbol{w}$ is an integration window with a limited extent towards the past, and $\boldsymbol{r_t(\tau)}$ is the autocorrelation function of lag $\boldsymbol{\tau}$ calculated at time $\boldsymbol{t}$. Licklider suggested that it could be calculated using synaptic or conduction delays (lag), coincidence-counting neurons (multiplication) and postsynaptic temporal summation (integration).

A fourth idea is to replace the multiplication operation of cross- or autocorrelation by subtraction (implemented by replacing excitatory by inhibitory neural interaction). This idea was used in the equalization-cancellation (EC) model of Durlach (1963) to explain binaural masking release, and the harmonic cancellation model of de Cheveign (1993) to explain segregation on the basis of harmonic structure. Cancellation is closely related to correlation.

Other ideas have been proposed, and still more are no doubt lurking in ingenious minds. Patterson (1987) proposed a "pulse-ribbon" or "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 pattern (rather than autocorrelating the latter). Cariani (2001) describes recurrent networks within which patterns circulate and are correlated (or convolved) with incoming patterns. More generally, Maass (1998) has shown that networks of spiking neurons are at least as powerful as (and in some ways more than) formal neural networks of sigmoidal neurons, time-of-arrival replacing average rate. In particular they can implement filters with arbitrary transfer functions. While this idea has mostly been explored in modalities other than hearing, it certainly makes sense to try it out in this very temporal modality. The next section describes a few models that utilize some of these ideas for specific tasks.

3. Models that use time

3.1 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 autocorrelation.

The autocorrelation model accounts for a wide range of pitch phenomena (Meddis and Hewitt, 1991; Cariani and 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 and Srulovicz, 1977), is that it does not require a 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 and Shackleton, 1994). This is currently a major issue in hearing. From an applied point of view, autocorrelation has proved to be an effective basis for fundamental frequency estimation of speech and music (de Cheveign and Kawahara, 2002).

The amount of time used to calculate the running autocorrelation function (Eq. 1) is a sum of two terms. The first is the range of values of $\boldsymbol{\tau}$ that is explored, the second the size of the integration window $\boldsymbol{w}$. 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 of at least $2T_{MAX}$. Actually it is possible to reduce this to $T+T_{MAX}$, where T is the period being measured (de Cheveign and Kawahara, 2002). It is *not* possible to go below this. A pitch model that uses a shorter-than-T_{MAX} integration window is not complete, in that it doesn’t explain how its fluctuating behaviour produces a stable percept of pitch.

Meddis and Hewitt (1991) used an exponential window of time constant 2.5 ms, adequate for fundamental frequencies beyond 400 Hz, but Meddis and Hewitt (1992) later proposed a longer window of 10 ms. Recent efforts to determine the appropriate size experimentally were reviewed by Wiegrebe (2001). It turns out that the integration window size is task dependent, and in particular may vary with F0. Data are consistent with an integration duration of twice the period with a minimum duration of 2.5ms. The only problem with this proposition is that period-dependent integration assumes prior knowledge of the period, a circular process.

Integration can cover longer windows (Moore, 1973), and it can be "reset" by transient events (Plack and White, 2000a), as has also been observed in the binaural system (Hafter and Buel, 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, this workshop). This idea is closely related to that of missing features in automatic speech recognition (Cooke et al., 1997). As another way to improve resolution, it has been proposed that higher-order peaks of the autocorrelation function (at 2T, 3T, etc.) might be used in addition to the peak at T (Yost, 1999; Plack and White, 2000b; de Cheveign , 2000a). This also requires time. For a given stimulus duration there are various possible tradeoffs between lag range and integration duration, so the two factors are confounded in experiments that attempt to probe them.

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 observed for stimuli with resolved v.s. unresolved components. The minimum integration time depends upon the period, but integration windows may be longer depending on the task. Functionally, longer windows favour accuracy while shorter windows allow fast modulations to be followed (or at least recognized as pitch-like).

1.2 Timbre

The autocorrelation function has also been used to account for the perception of timbre. Meddis and Hewitt (1992) used template matching of the low-lag portion of the autocorrelation (below 2.5-4ms) in a model of vowel identification. The motivation for a time-domain model is weaker than in the case of pitch, as peripheral selectivity is sufficiently fine (if not too fine) to resolve formant patterns.

A possible functional advantage of an autocorrelation-based mechanism is that it allows F0-adaptive truncation, equivalent to sampling the spectrum precisely at harmonics of F0 (or calculating the Fourier transform over exactly one period) (Kawahara et al., 1999; de Cheveign and Kawahara, 1999). This goes some way to solving the old problem of F0-dependency of estimates of spectral shape (Klatt, 1982). The autocorrelation function (equivalent to the *power* spectrum) is over-sensitive to high-amplitude spectral features, one reason why the cepstrum (Fourier transform of the *log* spectrum) is preferred for speech processing. However calculating ACFs within independent channels after amplitude normalization has the similar effect of producing a well-balanced representation.

To summarize, a model exists to 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.

1.3 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 and Carlyon, 1995) *harmonic structure* is useful when competing voices have different fundamental frequencies (Brokx and Nooteboom, 1982). Among the many models proposed, those based on a spectral representation don’t work well if given a resolution consistent with that of the cochlea (Parsons, 1976; Assmann and Summerfield, 1990; de Cheveign , 1993). This is a task for which temporal processing seems necessary.

Time-domain models work according to either of two principles. With *channel selection*, the temporal pattern within each channel is used to assign it to one source or another (Meddis and Hewitt, 1992). With *channel splitting*, channels are shared between sources (de Cheveign , 1993, 1997). Channel selection relies on peripheral filtering for an initial analysis, channel splitting does not (although the initial filtering may make things easier by improving signal-to-noise ratio). Channel selection fails if all channels are dominated by interference, but experiments show that segregation still occurs in that situation, so channel selection cannot provide a complete account. A likely proposition is that both principles are at work. Temporal patterns were autocorrelated in Meddis and Hewitt’s model, transformed to a modulation spectrum in that of Meyer and Berthommier (1996), or filtered by a "neural cancellation" filter (followed by autocorrelation) in the model of de Cheveign (1993, 1997), and yet other schemes have been proposed (Assmann and Summerfield, 1990; Brown et al. 1996; Cooke and Ellis, 2001).

Assmann and Summerfield (1994) found that pairs of concurrent vowels with different F0s were harder to segregate if shortened from 200 to 50 ms. McKeown and Patterson (1995) found that one vowel in each pair was usually dominant, and could be identified from as little as one cycle (at a 100 or 200 Hz fundamental), while identification of the second improved gradually with duration up to 8 cycles. This suggests a certain "sluggishness" of the segregation mechanism. On the other hand, Culling et al. (1994) found little difference according to whether F0s 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 Cheveign (1993; 1997) both use the F0 of the dominant vowel to retrieve the weaker vowel. Longer stimuli may help to estimate this F0 in the presence of the other vowel.

To summarize, harmonicity-based segregation appears to be based on temporal processing. Segregation probably involves estimating the fundamental frequency of an interfering voice to retrieve the target voice, a process that takes time.

1.4 Segregation based on binaural cues

As for harmonicity, binaural segregation models exist that use both principles of *channel selection* and *channel splitting*. As an example of the first, Lyon (1983) calculated cross-correlation functions individually within each channel, and grouped those that had a peak at the internal delay that matched the external interaural delay of a source. An example of the second is the equalization-cancellation (EC) model of Durlach (1963), that has recently been revisited by Culling and Summerfield (1995) and Breebart et al. (2001). The EC model adjusts (internally) the relative amplitude and delay of signals from both 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 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 produce high MLDs do not always produce high intelligibility level differences (ILDs) for speech. The benefit of spatial cues is not simply the result of instantaneous unmasking, but appears also to involve organization of speech parts across time (Darwin and Hukin, 1999). This process is not yet well understood.

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

2. Discussion and Conclusion

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. Spectrotemporal resolution is not necessarily determined by that of the cochlea. Does this mean that spectrotemporal excitation patterns (Moore, this workshop) are wrong and should be replaced? The answer is no for several reasons. One is that those patterns have good predictive power for a wide range of phenomena. It may be that temporal processing has properties that map well to a description in terms of cochlear excitation patterns. Another is that 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) are rich representations, but partly for that reason they are not as convenient as spectrotemporal patterns. A wise course may be to stick with the latter as a descriptive tool, while remaining alert for phenomena that escape them.

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 aim was to give insight into the sort of functions that can be purveyed, and the palette of mechanisms that might purvey them. A more authoritative account is way into the future.

A third purpose was to give some indication of time constants, on the basis of functional and/or behavioural considerations. Time is required to measure the scale 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 longer durations when available, and to adaptively tailor information to maximize performance.

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 constrains their order. All these aspects must be taken into account in the design of models to process time.

Acknowledgements

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," Aud. Neurosci. 3, 335-350.

Akeroyd, M. A., and Bernstein, L. R. (2001). "The variation across time of sensitivity to interaural disparities: Behavioral measurements and quantitative analyses," J. Acoust. Soc. Am. 110, 2516-2526.

Assmann, P. F., and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. 88, 680-697.

Assmann, P. F., and Summerfield, Q. (1994). "The contribution of waveform interactions to the perception of concurrent vowels," J. Acoust. Soc. Am. 95, 471-484.

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

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

Brown, G. J., Cooke, M., and Mousset, E. (1996). "Are neural oscillations the substrate of auditory grouping?", Proc. Workshop on the auditory basis of speech perception, Keele, 174-179.

Cariani, P. (2001). "Neural timing nets for auditory computation," in "Computational models of auditory function," Edited by S. Greenberg and M. Slaney, Amsterdam, IOS Press, 233-247.

Cariani, P. A., and Delgutte, B. (1996). "Neural correlates of the pitch of complex tones. I. Pitch and pitch salience," J. Neurophysiol. 76, 1698-1716.

Carlyon, R. P., and Shackleton, T. M. (1994). "Comparing the fundamental frequencies of resolved and unresolved harmonics: evidence for two pitch mechanisms?," J. Acoust. Soc. Am. 95, 3541-3554.

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

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

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

Culling, J. F., and Summerfield, Q. (1995). "Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. 98, 785-797.

Darwin, C. J., and Carlyon, R. P. (1995). "Auditory grouping," in "Handbook of perception and cognition: Hearing," Edited by B. C. J. Moore, New York, Academic Press, 387-424.

Darwin, C. J., and 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., Püschel, D., and Kohlrausch, A. (1996). "A quantitative model of the ‘‘effective’’ signal processing in the auditory system. I. Model structure," J. Acoust. Soc. Am. 99, 3615-3622.

Dau, T., and Kollmeier, B. (1997). "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration.," J. Acoust. Soc. Am. 102, 2906-2919.

de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. 93, 3271-3290.

de Cheveigné, A. (1997). "Concurrent vowel identification III: A neural model of harmonic interference cancellation," J. Acoust. Soc. Am. 101, 2857-2865.

de Cheveigné, A., and Kawahara, H. (1999). "Missing data model of vowel perception," J. Acoust. Soc. Am. 105, 3497-3508.

de Cheveigné, A. (2000a). "A model of the perceptual asymmetry between peaks and troughs of frequency modulation," J. Acoust. Soc. Am. 107, 2645-2656.

de Cheveigné, A. (2000b), "Modèles de traitement auditif dans le domaine temps," Université Paris 6, unpublished habilitation dissertation.

de Cheveigné, A., and Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., in press.

Delgutte, B., and Kiang, N. Y.-S. (1984a). "Speech coding in the auditory nerve: I. Vowel-like sounds," J. Acoust. Soc. Am. 75, 866-878.

Delgutte, B., and Kiang, N. Y.-S. (1984b). "Speech coding in the auditory nerve: V. Vowels in background noise," J. Acoust. Soc. Am. 75, 908-918.

Delgutte, B. (1984). "Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds," J. Acoust. Soc. Am. 75, 879-886.

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

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

Goldstein, J. L., and Srulovicz, P. (1977). "Auditory-nerve spike intervals as an adequate basis for aural frequency measurement," in "Psychophysics and physiology of hearing," Edited by E. F. Evans and J. P. Wilson, London, Academic Press, 337-347.

Hafter, E. R., and Buell, T. N. (1990). "Restarting the adapted binaural system," J. Acoust. Soc. Am. 88, 806-812.

Jeffress, L. A. (1948). "A place theory of sound localization," J. Comp. Physiol. Psychol. 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.," J. Neurophysiol. 76, 2137-2156.

Kawahara, H., Masuda-Katsuse, I., and 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 technical report, MIT Research Monograph 35.

Klatt, D. H. (1982). "Speech processing strategies based on auditory models," in "The representation of speech in the peripheral auditory system," Edited by R. Carlson and B. Granström, Amsterdam, Elsevier, 181-196.

Kollmeier, B., and Gilkey, R. (1990). "Binaural forward and backward masking: evidence for sluggishness in binaural detection," J. Acoust. Soc. Am. 87, 1709-1719.

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

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

Licklider, J. C. R. (1959). "Three auditory theories," in "Psychology, a study of a science," Edited by S. Koch, New York, McGraw-Hill, I, 41-144.

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

Maass, W. (1998). "On the role of time and space in neural computation," Lecture notes in computer science 1450, 72-83.

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

McKeown, J. D., and Patterson, R. D. (1996). "The time course of auditory segregation: Concurrent vowels that vary in duration," J. Acoust. Soc. Am. 98, 1866-1877.

Moore, B. C. J. (1973). "Frequency difference limens for short-duration tones," J. Acoust. Soc. Am. 54, 610-619.

Moore, B. C. J. (2002). "Temporal integration and context effects in hearing.", this workshop.

Oertel, D. (1999). "The role of timing in the brain stem auditory nuclei of vertebrates," Ann. Rev. Physiol. 61, 497-519.

Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. 60, 911-918.

Patterson, R. D., Ed. (1987). "A pulse-ribbon model of peripheral auditory processing,".

Patterson, R., Anderson, T. R., and Francis, K. (1996). "Binaural auditory images and a noise-resistant, binaural auditory spectrogram for speech recognition.", Proc. Workshop on the auditory basis of speech perception, Keele, 245-252.

Peterson, G. E., and Barney, H. L. (1952). "Control methods in a study of the vowels," J. Acoust. Soc. Am. 24, 175-184.

Plack, C. J., and White, L. J. (2000a). "Perceived continuity and pitch perception," J. Acoust. Soc. Am. 108, 1162-1169.

Plack, C. J., and White, L. J. (2000b). "Pitch matches between unresolved complex tones differing by a single interpulse interval," J. Acoust. Soc. Am. 108, 696-705.

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

Robinson, K., and Patterson, R. D. (1995). "The stimulus duration required to identify vowels, their octave, and their pitch chroma," J. Acoust. Soc. Am. 98, 1858-1865.

Rutherford, W. (1886). "A new theory of hearing," J. Anal. Physiol. 21, 166-168.

Sabatini, B. L., and Regehr, W. G. (1999). "Timing of synaptic transmission," Ann. Rev. Physiol. 61, 521-542.

Schwartz, I. R. (1992). "The superior olivary complex and lateral lemniscal nuclei," in "The mammalian auditory pathway: neuroanatomy," Edited by D. B. Webster, A. N. Popper and R. R. Fay, New York, Springer-Verlag, 117-167.

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," J. Acoust. Soc. Am. 78, 1622-1632.

Shamma, S., and Klein, D. (2000). "The case of the missing pitch templates: how harmonic templates emerge in the early auditory system," J. Acoust. Soc. Am. 107, 2631-2644.

Wever, E. G., and 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," J. Acoust. Soc. Am. 109, 1082-1091.

Yost, W. A. (1999). "Pitch-strength discrimination involving regular interval stimuli," ARO abstract #232.

Young, E. D., and Sachs, M. B. (1979). "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," J. Acoust. Soc. Am. 66, 1381-1403.

Grosjean, F., and Gee, J.P. (1987) Prosodic structure and spoken word recognition. In U.H. Frauenfelder and L.K. Tyler (eds.), *Spoken Word Recognition.* Cambridge, MA: MIT Press. 135-155.