Coincidence Detection in Pitch Perception

S. Shamma, D. Klein and D. Depireux

Electrical and Computer Engineering Department, Institute for Systems Research
University of Maryland at College Park, College Park, MD 20742, USA
sas@eng.umd.edu

1. Introduction

Pitch plays a critical role in the perception of speech prosody, melody of music, and in organizing the acoustic environment into different sources. Like timbre, pitch refers to many distinct percepts with a host of confusing terms. They include "spectral pitch" evoked by sinusoidal signals, "residue pitch" associated with unresolved (high) harmonics, very slow click trains, and envelope of amplitude-modulated noise and sinusoids, and "periodicity pitch" (also known as virtual and missing fundamental pitch) evoked by low order, spectrally resolved harmonic tone complexes.

There is general agreement on the distinctive properties of the latter two percepts. For instance, periodicity pitch is more salient than residue pitch, has lower JND's, and covers a wider frequency range (up to 2000 Hz *versus* 400 Hz). It is also largely insensitive to the phases of the harmonic complex, and its saliency increases proportional to the number of resolved harmonic components (especially the 2nd-5th harmonics, known as the *dominance region*).

2. Spectral versus temporal algorithms

It is uncertain what the physiological bases of these pitch percepts are, and whether unitary or multiple mechanisms are involved. The uncertainty partially stems from the fact that tone complexes can be neurally encoded both as spectral patterns (due to cochlear analysis) and as temporal patterns (due to phase-locking to the individual components, or to the envelopes of their interactions). Each of these representations suggests radically different computations and mechanisms to extract the pitch. For instance, "spectral pitch theories" (Goldstein 1973) propose that pitch value is derived (centrally) from a comparison between an input spectral profile (regardless of how this profile is computed) and internally stored spectral templates consisting of the harmonic series of all possible fundamentals. Such algorithms are able to explain much about periodicity pitch but are incapable of accounting for residue pitch and hence one must postulate different mechanisms for its computation.

The absence of a ``unitary pitch mechanism", and of physiological and anatomical evidence in support of the harmonic templates, has given strong impetus to alternative "temporal" algorithms that can account for both percepts. These utilize the temporal response patterns in each auditory channel, and then combine information across all channels to get the final pitch estimate. The algorithms vary enormously in the nature of

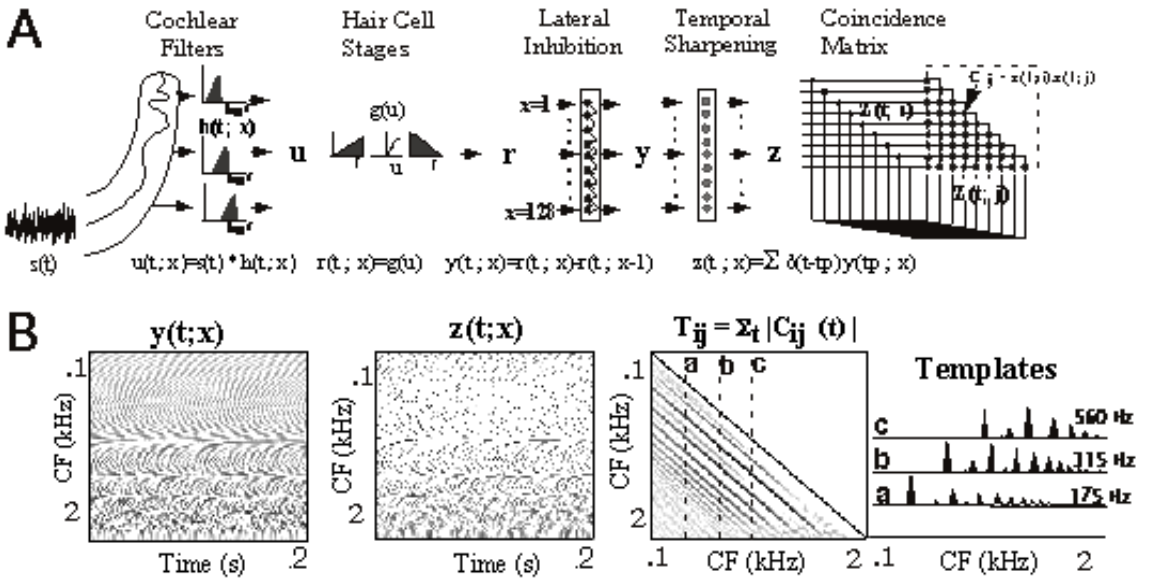

Figure 1 *(A) A schematic of early auditory processing and coincidence detection. (B) Representative responses to white noise at different stages of the model. The coincidence matrix output shown is the long-term accumulated response.*

cues and mechanisms they employ, e.g., first- or higher-order intervals (Cariani 1996), autocorrelations of the responses (Slaney and Lyon 1993, Meddis and Hewitt 1991, de Cheveigne 1998), or synchronization measures and oscillators (Langner and Schreiner 1988). A distinguishing characteristic of the temporal algorithms is that they make no use of a spectral pattern as such, and hence tonotopic order plays no role in their function.

However, the physiological basis of these models is also uncertain since data exhibiting delays or oscillatory responses do not coalesce as a whole into a compelling picture. Furthermore, most physiological pitch data tend to be in frequency ranges and from units with best frequencies that are more relevant for residue pitch (> 4 kHz) or slow temporal modulations (< 30 Hz) (Schreiner and Langner 1988) rather than periodicity pitch. Finally, recent psychoacoustical findings have been interpreted in favor of a dual (rather than a unitary) model of pitch perception (Carlyon 1998).

We describe here a *unitary algorithm* for pitch estimation that is ``spectral'' in character, i.e., it *does not* require neural delay-lines for correlators, oscillators, and other ``temporal" mechanisms. Instead, it is inspired by a *coincidence* mechanism that effectively gives rise to the harmonic templates – and hence to periodicity pitch – and, simultaneously, to a representation of residue pitch where appropriate. We shall first review the model, and the physiological and anatomical support for its realization. We then formulate the computational algorithm and illustrate its results. A more detailed account and discussion of the model is available in (Shamma and Klein 2000).

3. The coincidence matrix model

The model assumes two basic processing stages as illustrated in Figure 1. The first is an analysis stage consisting of the cochlear filter bank followed by temporal and spectral sharpening analogous to the processing seen in the cochlea and cochlear nucleus.

The second stage is a matrix of coincidence detectors that computes the pair-wise instantaneous correlation among all filter outputs. The filter-bank consists of 128 band-pass filters with a constant Q and equally spaced center frequencies (CFs) along 5.3 octaves of a logarithmic frequency axis *x.* The input signal *s(t)* is convolved with the filter impulse responses *h(t;x)* to produce the basilar membrane displacement *u(t;x)*. A hair cell stage includes (optional) filtering and a non-linearity, *g(.),* leading to auditory-nerve response rate *r(t;x)*. Next, the responses are sharpened spectrally by a lateral inhibitory network modeled by a simple first-order derivative across the channel array (Shamma 1984, Rhode and Greenberg 1994). Finally, a temporal sharpening stage enhances the synchrony of the phase-locked responses, mimicking the transformations seen between the auditory-nerve and the onset units of the cochlear nucleus (Palmer et al. 1995). It is approximated by sampling the *positive peaks* of *y(t;x)* to produce the train of impulses $z(t;x)$ ($= \Sigma\, \delta(t-t_p)\, y(t_p;x)$), where t_p = locations of the positive peaks in time, and $\delta(.)$ is the discrete Dirac delta-function ($\delta(0)= 1$, and $\delta(.)=0$ otherwise).

The second coincidence stage performs an instantaneous match (e.g., product) between the responses *z(t)* of all pairs of channels (*i,j*) in the array to generate $C_{ij}\,(t)$, and then integrates the absolute values of all products over time to produce an adequately smoothed output $T_{ij.}$

4. Emergence of the harmonic templates

Figure 1B illustrates the response patterns at different stages, and the average pattern of coincidences T_{ij} that emerges in the lower CF regions (phase-locking region with CF < 2-3 kHz) when the coincidence network is stimulated by broadband noise. These strong coincidences form a pattern of multiple diagonals that are spaced at exactly harmonic intervals apart. Vertical sections across these patterns display peaks at CFs that are integral multiples of a fundamental frequency, i.e., can be thought of as the harmonic templates. These templates exhibit many of the desirable properties postulated in previous spectral pitch models such as the gradual de-emphasize of the higher harmonics.

Three factors contribute to the emergence of this coincidence pattern: (1*) Cochlear spectral resolution* segregates the phase-locked responses onto different channels (creating the tonotopic axis). (2*) Subsequent nonlinear half-wave rectification and enhancement of the temporal response* of each channel ensures that non-zero correlations occur between harmonically related segregated responses. To appreciate this argument, consider a 200 and a 400 Hz phase-locked response on two channels. If the responses are purely sinusoidal, then they will be orthogonal and the inner-product will always be zero (it will only be non-zero if the two channels contain the same frequency). If however, the response waveforms are nonlinearly distorted (e.g., half-wave rectified), each will effectively contain ``distortion" components at multiples of the original frequency. For example, the 200 Hz response will become a harmonic series containing some energy at 400 Hz, in addition to other higher multiples. In this case, the inner-product between the distorted 200 and 400 Hz responses may be non-zero. (3) Finally*, the rapid phase-shifts of the traveling wave* near its resonance guarantee that the phase-locked responses at a given CF are available locally in various phases. Consequently, there are always some channels that produce non-zero correlations with responses at harmonically related CFs *regardless* of the absolute phase of the stimulus components.

Since all three factors contribute incrementally to the function of the model, the exact parameters of each do not critically affect the results. Thus, decreasing phase-locking, broadening the filter bandwidths, or diminishing temporal synchrony and a softer rectification, all *gradually* abolish the clarity of the coincidence peaks.

5. Residue pitch patterns

Many signals evoke cochlear responses that are highly synchronized *across* the auditory-nerve array, especially in the higher CF regions, as shown in Figure 2A (left panel). Such signals include *unresolved* harmonic complexes with a large number of in-phase components (slow click trains < 50 Hz) or slowly amplitude modulated tones or noise. The coincidence patterns evoked consist of stripes that parallel the main diagonal (Fig.2A, right panel). The stripes occur at a distance (D) from the diagonal that is inversely proportional to the *repetition period* of the auditory-nerve response (Fig.2A), which is usually equal to the residue pitch perceived for these stimuli.

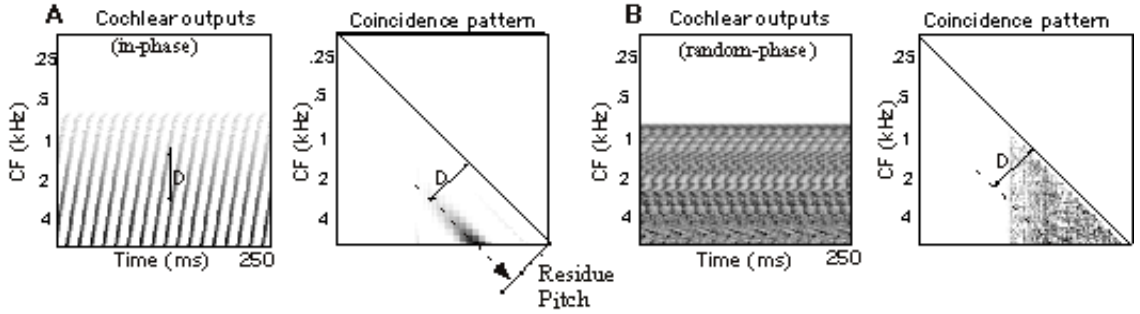

Figure 2. *(A) Cochlear responses to 10-110 unresolved harmonics of a 70 Hz fundamental and the corresponding coincidence patterns. The slight ŅtiltÓ in the responses is due to progressively increasing latency towards the lower CF channels (Shamma and Klein 2000). (B) Responses and coincidence pattern to the random-phase version of the stimulus above.*

This representation of residue pitch reflects many of its well-known perceptual properties. For instance, unlike periodicity pitch, residue pitch is sensitive to the phase of the stimulus components, becoming less salient as the phases are randomized. In the model, randomizing the stimulus phases destroys the response synchrony and the coincidence stripe (Fig.2B).

6. Physiological realizations of the model

The spectral and temporal analysis and sharpening in the first stage of the model have relatively unambiguous physiological correlates in the responses of the auditory nerve and cochlear nucleus cells.

The biological bases of the coincidence stage are less clear. Figure 3 illustrates two plausible implementations. The first interpretation (Fig.3A) is particularly suitable for the formation of the harmonic templates. Each neuron here has an extensive dendritic tree with synapses that correspond to a whole column of coincidence detectors. The synapses, which are initially weak, begin to strengthen during the learning phase at CF locations where synaptic inputs correlate well with the primary somatic input. In the end, each cell will be driven at many harmonically-related CFs, and hence will appear broadly tuned, but selective to a particular pitch value. Note that once the synapses are fully

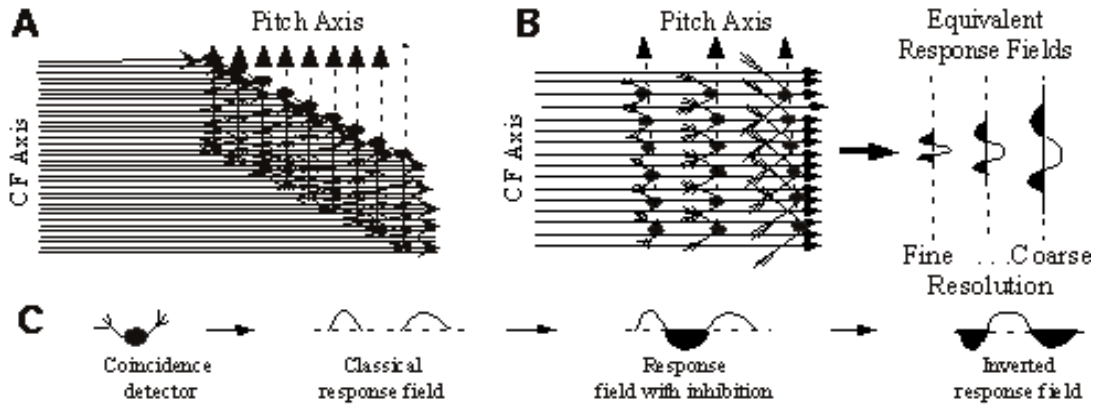

Figure 3 *(A,B) Two possible physiological realizations of the coincidence matrix model (see text for details). (C) How a lateral inhibitory response field may be viewed as a coincidence detetctor*

formed at the end of the learning phase, an input signal impinging at a strong synapse is capable of driving effectively the post-synaptic cell regardless of any other inputs.

Figure 3B depicts a second more literal interpretation of the model where each cell corresponds to a coincidence node in the matrix. This topology is more appropriate for computing the residue pitch, and in fact suggests an interesting relation between the coincidence matrix and multiscale representations thought to exist at the auditory cortex (Shamma 1996). This relationship is clarified in the sequence of transformations in Figure 3C where the meaning of a "coincidence detector" is redefined. For instance, the classical interpretation of a coincidence detector is of a neuron with a response field (RF) that sums (or multiplies) two inputs a certain CF distance apart. If one adds inhibition between the two excitatory fields and then inverts the sign of the output, one obtains the classic "lateral inhibitory" RF. Therefore, apart from an output inversion, the coincidence matrix can be physiologically realized as tonotopically organized layers of neurons (*corresponding to the diagonals* of the coincidence matrix) with lateral inhibitory RFs that gradually increase in bandwidth away from the main diagonal (Fig.4B, right panel).

This type of network produces a "multiscale" or "multiresolution" representation of its input pattern in the sense that if a spectral pattern is applied to the network, each layer transforms this input into a pattern that reflects the spectral features that are commensurate with its RF bandwidth. Thus, fine peaks and edges are prominently represented in the output of the narrowly-tuned or *fine resolution* RFs, while coarse spectral trends are depicted by the broadly tuned or *coarse resolution* RFs (Fig.3B). Similarly, for spectral patterns evoking residue pitch (as in Fig.2A), the most active layer reflects the CF distance D (or *resolution*) between the peaks, and hence the repetition rate of the stimulus.

7. Computing periodicity and residue pitch

The physiological interpretations in Figure 3 suggest specific algorithms to compute pitch values and saliency. These are schematically depicted in Figure 4. For periodicity pitch, the algorithm is based on Fig.3A, where each neuron along the "pitch

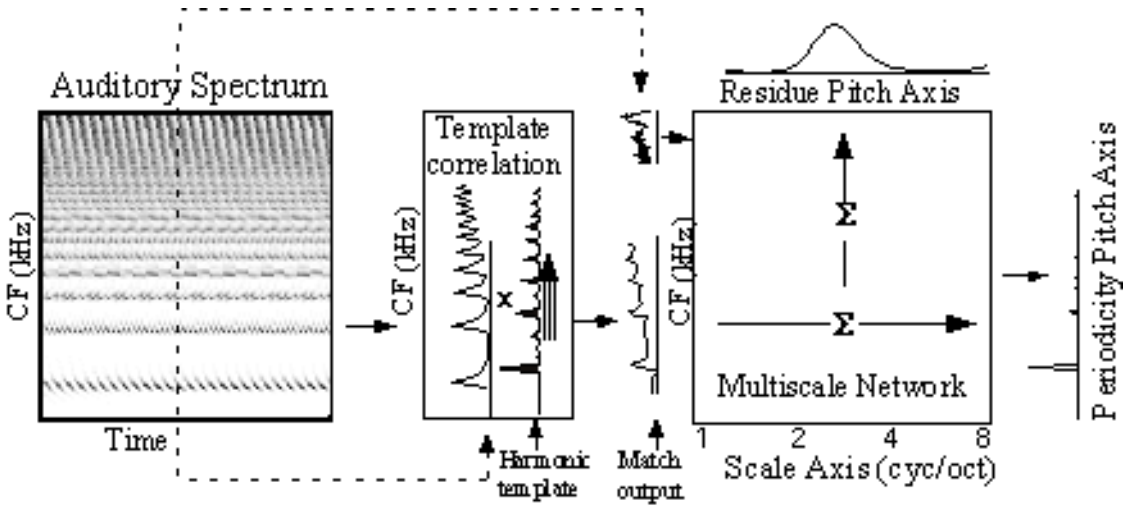

Figure 4 *Schematic of algorithm to extract periodicity and residue pitch. For periodicity pitch, the cross-section of the auditory spectrum at each instant is correlated with harmonic templates, and the result is enhanced by the multiscale network (which functions effectively as a coincidence matrix). For residue pitch, the, cross-section is applied directly to the multiscale network..*

axis" computes a weighted sum of inputs from harmonically related CFs.. This is equivalent to a correlation between a typical template (e.g., one cross section as in Fig.1B) and the incoming spectrum, much like the classical spectral pitch algorithms (Duifhuis 1982). Finally the output of the template match is applied to the multiscale network and the 2-dimensional result is collapsed across the *upper scales* only (above 1 cycle/octave) as shown in Figure 4. The multiscale network sharpens the matching output by removing its lower scales, or equivalently, its overall gradual trends.

Residue pitch can be computed by summing the coincidence matrix output along the diagonals as illustrated by the dashed lines in Fig.2A (right panel). Alternatively, the spectrum can be applied to the multiscale network, and then sum its output within each scale as depicted in Fig.4. The resulting pattern exhibits peaks that are usually broader than those of periodicity pitch (larger JNDs?) and whose height depends directly on the bandwidth of the *synchronized* responses.

For both pitch estimation algorithms, we normalize and interpret the final result of the summation as a probability density function. The height of each peak then is taken to reflect the saliency of the perceived pitch. The results of such computations are illustrated in Fig.5A for two speech stimuli. Figure 5B demonstrates the important role of periodicity pitch plays in grouping and separating the partials belonging to simultaneous sources with different pitches by applying a harmonic sieve whose fundamental is computed by the algorithm depicted earlier in Figure 4. Audio samples of these and other analogous grouping phenomena (e.g, the perception of a mistuned harmonic in a tone complex) are available at www.isr.umd.edu/CAAR/pubs.html.

8. Conclusion

A simple coincidence detection network explains both the formation of harmonic templates for periodicity pitch perception, and the representation of residue pitch. The network performs an *instantaneous* comparison of responses *across* the tonotopic axis,

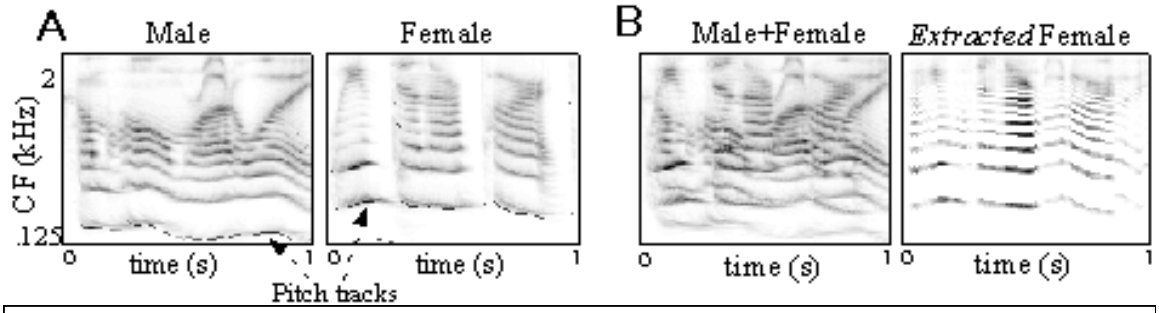

Figure 5 *(A) Periodicity pitch estimation for speech. (B) Separating speakers using harmonic sieve.*

and hence does not invoke any neural delays, oscillators, or other "temporal" operations. We have shown previously that this coincidence principle is useful in robustly extracting similar cues in very different contexts. These include *lateral inhibition* (essentially a form of coincidence as in Fig.3) to estimate a level-insensitive spectrum from saturated auditory-nerve responses (Shamma 1984), and the *stereausis* coincidence matrix to process interaural time-delays without requiring precise neural delay-lines (Shamma et al. 1989). Coincidence detection thus provides a unifying principle not only for pitch, but also for other auditory percepts, and even across sensory systems (deValois and deValois 1990).

9. Acknowledgements

This work is supported by ONR (ODDR\&E MURI97) and NSF (LIS-CMS9720334)

10. References

Cariani P. and Delgutte, B.(1996) Neural Correlates of the Pitch of Complex Tones. I: Pitch and Pitch Salience. Journal of Neurophysiology 76,1698-1716.

Carlyon R. (1998) Comments on a Unitary Model of Pitch. J. Acoust. Soc. Am. 104, 1118-1121.

de Cheveigne A. (1998) Cancellation Model of Pitch Perception. J. Acoust. Soc. Am. 103,1261-1271.

Duifhuis H. and Willems, L. and Sluyter, R. (1982) Measurement of Pitch in Speech: An implementation of Goldstein's theory of pitch perception. J. Acoust. Soc. Am 71, 1568-1580.

Goldstein J. (1973) An optimum processor theory for the central formation of pitch of complex tones. J. Acoust. Soc. Am 54, 1496—1516.

Langner G. and Schreiner, C (1988) Periodicity Coding in the Inferior Colliculus of the Cat. Journal of Neurophysiology. 60,1805-1822.

Meddis R. and J. Hewitt (1991) Virtual Pitch and Phase Sensitivity of a Computer Model of the Auditory Periphery. I: Pitch Identification. J. Acoust. Soc. Am. 89, 2866-2882.

Palmer A.,Winter, I., Jiang, D. and James, N.(1995) Across Frequency Integration by Neurons in the Ventral Cochlear Nucleus. In *Advances in Hearing Research* J. Manley and Klump, G. and Kopple, C. and Fastl, H. and Oeckinghaus,H. (Eds.)". World Scientific Publishers. Singapore

Rhode W. and S. Greenberg (1994) Lateral suppression and inhibition in the cochlear nucleus of the cat. J. Neurophysiol. 71,493-519.

Shamma S. (1984) 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. (1996) Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method. Network: Computation in Neural Systems 7, 439-476.

Shamma, S. N. Shen and P.Gopalaswamy (1989) Stereausis: Binaural processing without neural delays. J. Acoust. Soc. Am.86,989-1006

Shamma S. and D. Klein (2000) The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. J. Acoust. Soc. Am. (in press).

Slaney M. nd R. Lyon (1993) On the importance of time - A temporal representation of sound. In M. Cooke and S. Beet and M. Crawford (Eds.) *Visual Representations of Speech Signals*. J. Wiley and Sons, Sussex England.

de Valois R. and K. de Valois (1990) *Spatial Vision*. Oxford University Press.