Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform[a)]

Roy D. Patterson and Mike H. Allerhand
MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, United Kingdom

Christian Giguère
Laboratory of Experimental Audiology, University Hospital Utrecht, 3508 GA Utrecht, The Netherlands

(Received 8 December 1994; revised 5 April 1995; accepted 21 April 1995)

A software package with a modular architecture has been developed to support perceptual modeling of the fine-grain spectro-temporal information observed in the auditory nerve. The package contains both functional and physiological modules to simulate auditory spectral analysis, neural encoding, and temporal integration, including new forms of periodicity-sensitive temporal integration that generate stabilized auditory images. Combinations of the modules enable the user to approximate a wide variety of existing, time-domain, auditory models. Sequences of auditory images can be replayed to produce cartoons of auditory perceptions that illustrate the dynamic response of the auditory system to everyday sounds. © *1995 Acoustical Society of America.*

PACS numbers: 43.66.Ba, 43.64.Bt, 43.71.An

INTRODUCTION

Several years ago, we developed a functional model of the cochlea to simulate the phase-locked activity that complex sounds produce in the auditory nerve. The purpose was to investigate the role of fine-grain timing information in auditory perception, generally (Patterson *et al.*, 1992b; Patterson and Akeroyd, 1995), and in speech perception, in particular (Patterson *et al.*, 1992a). The architecture of the resulting auditory image model (AIM) is shown in the left-hand column of Fig. 1. The responses of the three modules to the vowel in "hat" are shown in the three panels of Fig. 2. Briefly, the spectral analysis stage converts the sound wave into the model's representation of basilar membrane motion (BMM). For the vowel in hat, each glottal cycle generates a version of the basic vowel structure in the BMM (top panel). The neural encoding stage stabilizes the BMM in level and sharpens features like vowel formants, to produce a simulation of the neural activity pattern (NAP) produced by the sound in the auditory nerve (middle panel). The temporal integration stage stabilizes the repeating structure in the NAP and produces a simulation of our perception of the vowel (bottom panel), referred to as the *auditory image*. Sequences of simulated images can be generated at regular intervals and replayed as an animated cartoon to show the dynamic behavior of the auditory images produced by everyday sounds.

An earlier version of the AIM software was made available to collaborators via the Internet. From there it spread to the speech and music communities, indicating a more general interest in auditory models than we had originally anticipated. This has prompted us to prepare documentation and a formal release of the software (AIM R7).

A number of users wanted to compare the outputs from the functional model, which is almost level *in*dependent, with those from physiological models of the cochlea, which are fundamentally level dependent. Others wanted to compare the auditory images produced by strobed temporal integration with correlograms. As a result, we have installed alternative modules for each of the three main stages as shown in the right-hand column of Fig. 1. The alternative spectral analysis module is a nonlinear, transmission line filterbank based on Giguère and Woodland (1994a). The neural encoding module is based on the inner haircell model of Meddis (1988). The temporal integration module generates correlograms like those of Slaney and Lyon (1990) or Meddis and Hewitt (1991), using the algorithm proposed by Allerhand and Patterson (1992). The responses of the three modules to the vowel in hat are shown in Fig. 3 for the case where the level of the vowel is 60 dB SPL. The patterns are broadly similar to those of the functional modules but the details differ, particularly at the output of the third stage. The differences grow more pronounced when the level of the vowel is reduced to 30 dB SPL or increased to 90 dB SPL. Figures 2 and 3 together illustrate how the software can be used to compare and contrast different auditory models. The new modules also open the way to time-domain simulation of hearing impairment and distortion products of cochlear origin.

Switches were installed to enable the user to shift from the functional to the physiological version of AIM at the output of each stage of the model. This architecture enables the system to implement other popular auditory models such as the gammatone-filterbank, Meddis-haircell, correlogram models proposed by Assmann and Summerfield (1990), Meddis and Hewitt (1991), and Brown and Cooke (1994). The remainder of this letter describes the integrated software package with emphasis on the functional and physiological routes, and on practical aspects of obtaining the software package.

I. THE AUDITORY IMAGE MODEL

A. The spectral analysis stage

Spectral analysis is performed by a bank of auditory filters which converts a digitized wave into an array of fil-

[a)]Instructions for acquiring the software package electronically are presented in Sec. II. This document refers to AIM R7 which is the first official release.

 0001-4966/95/98(4)/1890/5/$6.00

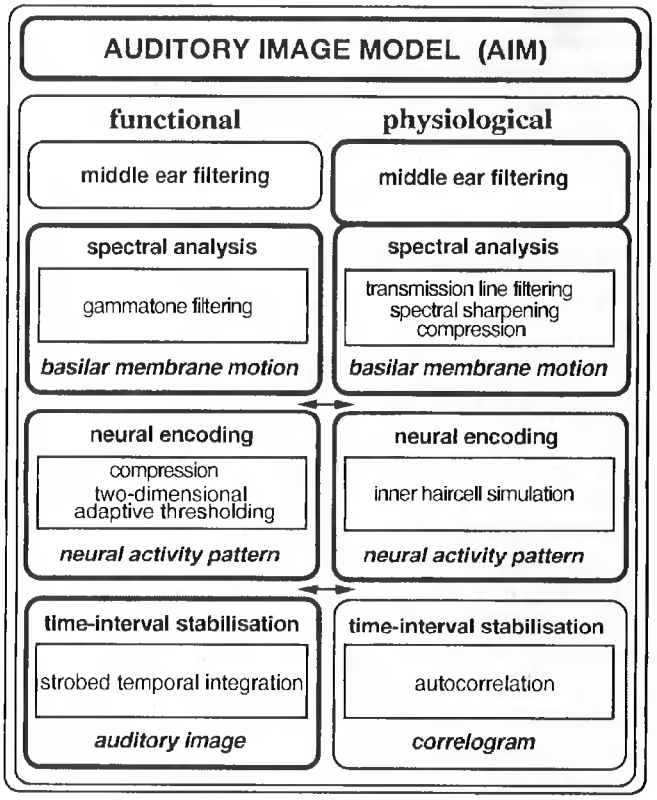

FIG. 1. The three-stage structure of the AIM software package. Left-hand column: functional route; right-hand column: physiological route. For each module, the figure shows the function (bold type), the implementation (in the rectangle), and the simulation it produces (italics).

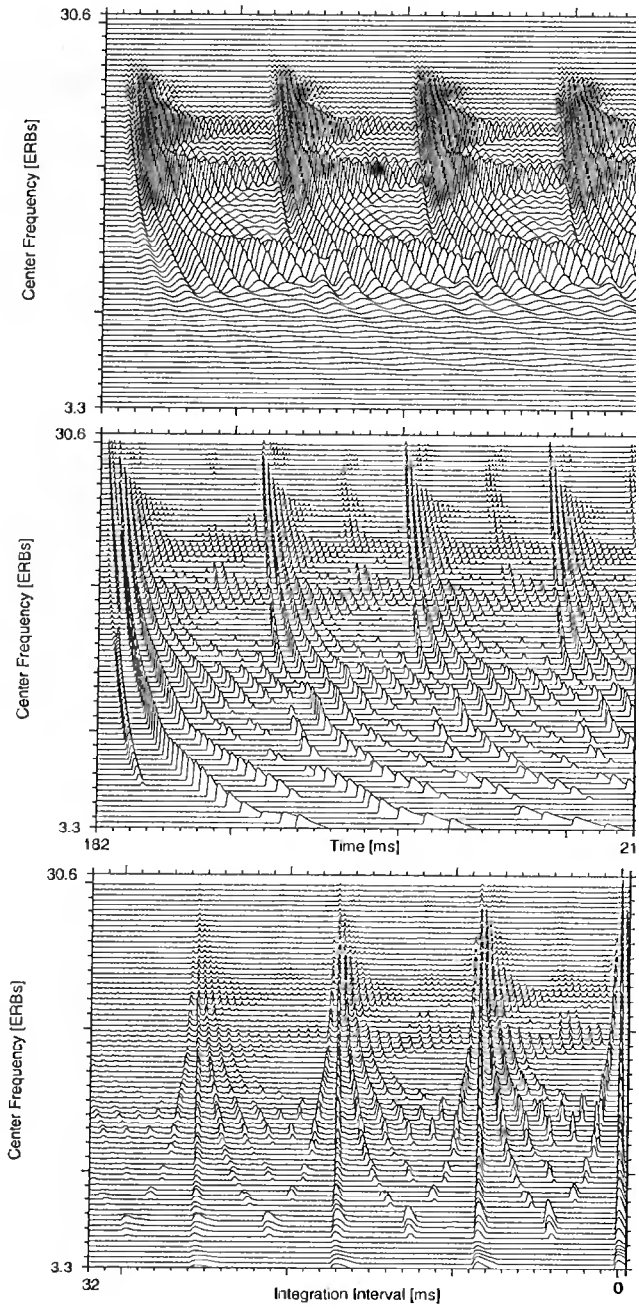

FIG. 2. Responses of the model to the vowel in hat processed through the functional route: (top) basilar membrane motion, (middle) neural activity pattern, and (bottom) auditory image.

tered waves like those shown in the top panels of Figs. 2 and 3. The set of waves is AIM's representation of basilar membrane motion. The software distributes the filters linearly along a frequency scale measured in equivalent rectangular bandwidths (ERBs). The ERB scale was proposed by Glasberg and Moore (1990) based on physiological research summarized in Greenwood (1990) and psychoacoustic research summarized in Patterson and Moore (1986). The constants of the ERB function can also be set to produce a reasonable approximation to the Bark scale. Options enable the user to specify the number of channels in the filterbank and the minimum and maximum filter center frequencies.

AIM provides both a functional auditory filter and a physiological auditory filter for generating the BMM: the former is a linear, gammatone filter (Patterson *et al.*, 1992b); the latter is a nonlinear, transmission line filter (Giguère and Woodland, 1994a). The impulse response of the gammatone filter provides an excellent fit to the impulse response of primary auditory neurons in cats, and its amplitude characteristic is very similar to that of the "roex" filter commonly used to represent the human auditory filter. The motivation for the gammatone filterbank and the available implementations are summarized in Patterson (1994a). The input wave is passed through an optional middle-ear filter adapted from Lutman and Martin (1979).

In the physiological version, a "wave digital filter" is used to implement the classical, one-dimensional, transmission line approximation to cochlear hydrodynamics. A feedback circuit representing the fast motile response of the outer haircells generates level-dependent basilar membrane motion (Giguère and Woodland, 1994a). The filterbank generates combination tones of the type $f_1 - n(f_2 - f_1)$ which propagate to the appropriate channel, and it has the potential to generate cochlear echoes. Options enable the user to customize the transmission line filter by specifying the feedback gain and saturation level of the outer haircell circuit. The middle ear filter forms an integral part of the simulation in

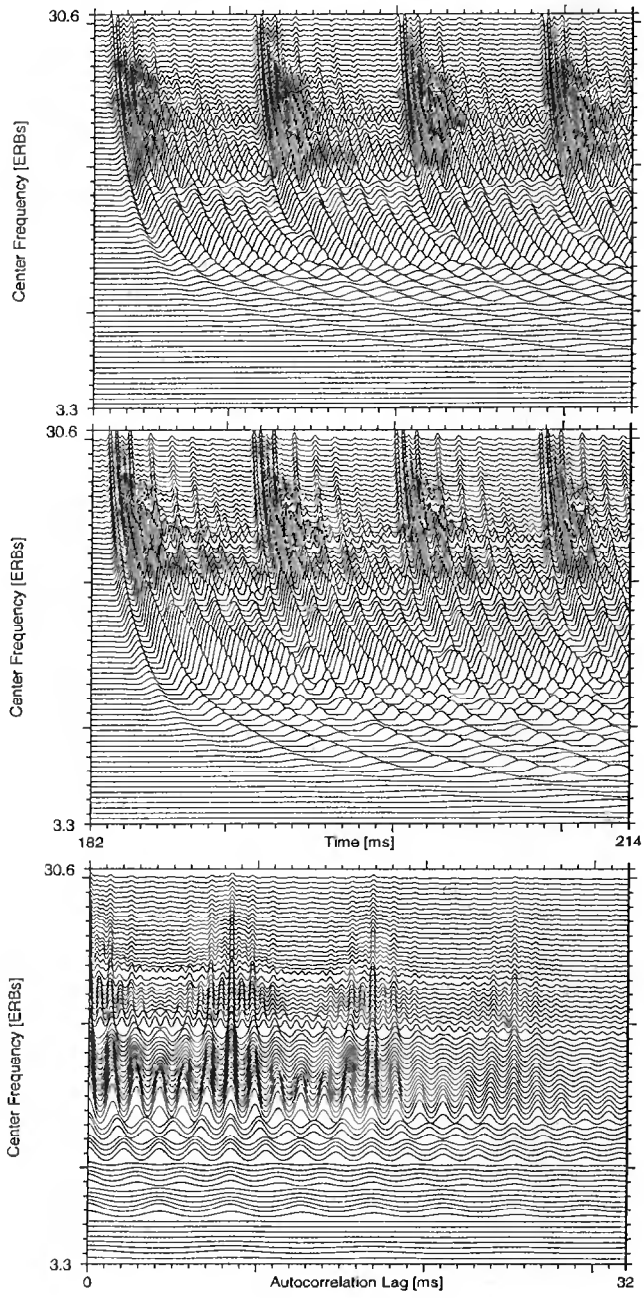

FIG. 3. Responses of the model to the vowel in hat processed through the physiological route: (top) basilar membrane motion, (middle) neural activity pattern, and (bottom) autocorrelogram image.

this case. Together, it and the transmission line filterbank provide a bidirectional model of auditory spectral analysis.

The upper panels of Figs. 2 and 3 show the responses of the two filterbanks to the vowel in hat. They have 75 channels covering the frequency range 100–6000 Hz (3.3–30.6 ERBs). In the high-frequency channels, the filters are broad and the glottal pulses generate impulse responses which decay relatively quickly. In the low-frequency channels, the filters are narrow and so they resolve individual continuous harmonics. The rightward skew in the low-frequency channels is the "phase lag," or "propagation delay," of the cochlea, which arises because the narrower low-frequency filters respond more slowly to input. The transmission line filterbank shows more ringing in the valleys than the gammatone filterbank because of its dynamic signal compression; as amplitude decreases the damping of the basilar membrane is reduced to increase sensitivity and frequency resolution.

B. The neural encoding stage

The second stage of AIM simulates the mechanical/neural transduction process performed by the inner haircells. It converts the BMM into a NAP, which is AIM's representation of the afferent activity in the auditory nerve. Two alternative simulations are provided for generating the NAP: a bank of two-dimensional adaptive thresholding units (Holdsworth and Patterson, 1991), or a bank of inner haircell simulators (Meddis, 1988).

The adaptive thresholding mechanism is a functional representation of neural encoding. It begins by rectifying and compressing the BMM; then it applies adaptation in time and suppression across frequency. The adaptation and suppression are coupled and they jointly sharpen features like vowel formants in the compressed BMM representation. Briefly, an adaptive threshold value is maintained for each channel and updated at the sampling rate. The new value is the largest of (a) the previous value reduced by a fast-acting temporal decay factor, (b) the previous value reduced by a longer-term temporal decay factor, (c) the adapted level in the channel immediately above, reduced by a frequency spread factor, or (d) the adapted level in the channel immediately below, reduced by the same frequency spread factor. The mechanism produces output whenever the input exceeds the adaptive threshold, and the output level is the difference between the input and the adaptive threshold. The parameters that control the spread of activity in time and frequency are options in AIM.

The Meddis (1988) module simulates the operation of an individual inner haircell; specifically, it simulates the flow of neurotransmitters across three reservoirs that are postulated to exist in and around the haircell. The module reproduces important properties of single afferent fibers such as two-component time adaptation and phase locking. The transmitter flow equations are solved using the wave-digital-filter algorithm described in Giguère and Woodland (1994a). There is one haircell simulator for each channel of the filterbank. Options allow the user to shift the entire rate-intensity function to a higher or lower level, and to specify the type of fiber (medium or high spontaneous rate).

The middle panels in Figs. 2 and 3 show the NAPs obtained with adaptive thresholding and the Meddis module in response to BMMs from the gammatone and transmission line filterbanks of Figs. 1 and 2, respectively. The phase lag of the BMM is preserved in the NAP. The positive half-cycles of the BMM waves have been sharpened in time, an effect which is more obvious in the adaptive thresholding NAP. Sharpening is also evident in the frequency dimension of the adaptive thresholding NAP. The individual "haircells" are not coupled across channels in the Meddis module, and thus there is no frequency sharpening in this case. The physi-

ological NAP reveals that the activity between glottal pulses in the high-frequency channels is due to the strong sixth harmonic in the first formant of the vowel.

C. The temporal integration stage

Periodic sounds give rise to static, rather than oscillating, perceptions indicating that temporal integration is applied to the NAP in the production of our initial perception of a sound—our auditory image. Traditionally, auditory temporal integration is represented by a simple leaky integration process and AIM provides a bank of low-pass filters to enable the user to generate auditory spectra (Patterson, 1994a) and auditory spectrograms (Patterson *et al.*, 1992a). However, the leaky integrator removes the phase-locked fine structure observed in the NAP, and this conflicts with perceptual data indicating that the fine structure plays an important role in determining sound quality and source identification (Patterson, 1994b; Patterson and Akeroyd, 1995). As a result, AIM includes two modules which preserve much of the time-interval information in the NAP during temporal integration, and which produce a better representation of our auditory images. In the functional version of AIM, this is accomplished with strobed temporal integration (Patterson *et al.*, 1992a, b); in the physiological version, it is accomplished with a bank of autocorrelators (Slaney and Lyon, 1990; Meddis and Hewitt, 1991).

In the case of strobed temporal integration (STI), a bank of delay lines is used to form a buffer store for the NAP, one delay line per channel, and as the NAP proceeds along the buffer it decays linearly with time, at about 2.5%/ms. Each channel of the buffer is assigned a strobe unit which monitors activity in that channel looking for local maxima in the stream of NAP pulses. When one is found, the unit initiates temporal integration in that channel; that is, it transfers a copy of the NAP at that instant to the corresponding channel of an image buffer and adds it point for point with whatever is already there. The local maximum itself is mapped to the 0-ms point in the image buffer. The multichannel version of this STI process produces AIM's representation of our auditory image of a sound. Periodic and quasiperiodic sounds cause regular strobing which leads to simulated auditory images that are static, or nearly static, and which have the same temporal resolution as the NAP. Dynamic sounds are represented as a sequence of auditory image frames. If the rate of change in a sound is not too rapid, as in diphthongs, features are seen to move smoothly as the sound proceeds, much as characters move smoothly in animated cartoons.

An alternative form of temporal integration is provided by the correlogram (Slaney and Lyon, 1990; Meddis and Hewitt, 1991). It extracts periodicity information and preserves intraperiod fine structure by autocorrelating each channel of the NAP. The correlogram is the multichannel version of this process. It was originally introduced as a model of pitch perception (Licklider, 1951) with a neural wiring diagram to illustrate that it was physiologically plausible. To date, however, there is no physiological evidence for autocorrelation in the auditory system, and the installation of the module in the physiological route was a matter of convenience. The current implementation is a recursive, or running, autocorrelation. A functionally equivalent FFT-based method is also provided (Allerhand and Patterson, 1992). A comparison of the correlogram in the bottom panel of Fig. 3 with the auditory image in the bottom panel of Fig. 2 shows that the vowel structure is more symmetric in the correlogram and there are larger level contrasts in the correlogram. It is not yet known whether one of the representations is more realistic or more useful. The present purpose is to note that the software package can be used to compare auditory representations in a way not previously possible.

II. THE SOFTWARE/HARDWARE PLATFORM

A. The software package

The code is distributed as a compressed archive (in unix tar format), and can be obtained via ftp from the address: ftp.mrc-apu.cam.ac.uk (Name=anonymous; Password =⟨your E-mail address⟩). All the software is contained in two archives: pub/aim/aimR7.tar.Z and pub/aim/aimR7docs.tar.Z. The associated text file pub/aim/ReadMe.First contains instructions for installing and compiling the software. The AIM package consists of a makefile and several subdirectories. Five of these (filter, glib, model, stitch, and wdf) contain the C code for AIM. An AIM/tools directory contains C code for ancillary software tools. These software tools are provided for pre/postprocessing of model input/output. A variety of functions are offered, including stimulus generation, signal processing, and data manipulation. An AIM/man directory contains on-line manual pages describing AIM and the software tools. An AIM/scripts directory contains demonstration scripts for a guided tour through the model. Sounds used to test and demonstrate the model are provided in the AIM/waves directory. These sounds were sampled at 20 kHz, and each sample is a 2-byte number in little-endian byte order; a tool is provided to swap byte order when necessary.

B. System requirements

The software is written in C. The code generated by the native C compilers included with ULTRIX (version 4.3a and above) and SUNOS (version 4.1.3 and above) has been extensively tested. The code from the GNU C compiler (version 2.5.7 and above) is also reliable. The total disk usage of the AIM source code is about 700 kbytes. The package also includes 500 kbytes of sources for ancillary software tools, and 200 kbytes of documentation. The executable programs occupy about 1000 kbytes, and executable programs for ancillary tools occupy 7000 kbytes. About 800 kbytes of temporary space are required for object files during compilation. The graphical interface uses X11 (R4 and above) with either the OPENWINDOWS or MOTIF user interface. The programs can be compiled using the base Xlib library (libX11.a), and will run on both 1-bit (mono) and multiplane (color or greyscale) displays.

C. Compilation and operation

The makefile includes targets to compile the source code for AIM and the associated tools on a range of machines

(DEC, SUN, SGI, HP); the targets differ only in the pathnames for the local X11 base library (libX11.a) and header files (X11/X.h and X11/Xlib.h). AIM can be compiled without the display code if the graphics interface is not required or if X11 is not available (make noplot). The executable for AIM is called gen. Compilation also generates symbolic links to gen, such as genbmm, gennap, and gensai, which are used to select the desired output (BMM, NAP, or SAI). The links and the executables for the AIM/tools are installed in the AIM/bin directory after compilation. Options are specified as name=value on the command line; unspecified options are assigned default values. The model output takes the form of binary data routed by default to the model's graphical displays. Output can also be routed to plotting hardware, or other postprocessing software.

III. APPLICATIONS AND SUMMARY

In hearing research, the functional version of AIM has been used to model phase perception (Patterson, 1987), octave perception (Patterson *et al.*, 1993), and timbre perception (Patterson, 1994b). The physiological version has been used to simulate cochlear hearing loss (Giguère *et al.*, 1993; Giguère and Woodland, 1994b), and combination tones of cochlear origin (Giguère *et al.*, 1995). In speech research, the functional version has been used to explain syllabic stress (Allerhand *et al.*, 1992), and both versions have been used as preprocessors for speech recognition systems (e.g., Patterson *et al.*, 1994; Giguère *et al.*, 1993).

In summary, the AIM software package provides a modular architecture for time-domain computational studies of peripheral auditory processing.

ACKNOWLEDGMENTS

The gammatone filterbank, adaptive thresholding, and much of the software platform were designed and written by John Holdsworth, the options handler is by Paul Manson, and the revised STI module by Jay Datta. Michael Akeroyd extended the postscript facilities and developed the xreview routine for auditory image cartoons. The software development was supported by grants from DRA Farnborough (U.K.), Esprit BR 3207 (EEC), and the Hearing Research Trust (U.K.). We thank Malcolm Slaney and Michael Akeroyd for helpful comments on an earlier version of the paper.

Allerhand, M., and Patterson, R. D. (**1992**). "Correlograms and auditory images," Proc. Inst. Acoust. **14**, 281–288.

Allerhand, M., Butterfield, S., Cutler, R., and Patterson, R. D. (**1992**). "Assessing syllable strength via an auditory model," Proc. Inst. Acoust. **14**, 297–304.

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.

Brown, G. J., and Cooke, M. (**1994**). "Computational auditory scene analysis," Comput. Speech Lang. **8**, 297–336.

Giguère, C., Woodland, P. C., and Robinson, A. J. (**1993**). "Application of an auditory model to the computer simulation of hearing impairment: Preliminary results," Can. Acoust. **21**, 135–136.

Giguère, C., and Woodland, P. C. (**1994a**). "A computational model of the auditory periphery for speech and hearing research. I. Ascending path," J. Acoust. Soc. Am. **95**, 331–342.

Giguère, C., and Woodland, P. C. (**1994b**). "A computational model of the auditory periphery for speech and hearing research. II: Descending paths," J. Acoust. Soc. Am. **95**, 343–349.

Giguère, C., Kunov, H., and Smoorenburg, G. F. (**1995**). "Computational modeling of psycho-acoustic combination tones and distortion-product otoacoustic emissions," 15th International Congress on Acoustics, Vol. III, pp. 237–240, Trondheim (Norway).

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–38.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.

Holdsworth, J. W., and Patterson, R. D. (**1993**). "Analysis of waveforms," UK Patent No. GB 2-234-078-B, UK Patent Office, London.

Licklider, J. C. R. (**1951**). "A duplex theory of pitch perception," Experientia **7**, 128–133.

Lutman, M. E., and Martin, A. M. (**1979**). "Development of an electroacoustic analogue model of the middle ear and acoustic reflex," J. Sound. Vib. **64**, 133–157.

Meddis, R. (**1988**). "Simulation of auditory-neural transduction: Further studies," J. Acoust. Soc. Am. **83**, 1056–1063.

Meddis, R., and Hewitt, M. J. (**1991**). "Modeling the perception of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. **91**, 233–45.

Patterson, R. D. (**1987**). "A pulse ribbon model of monaural phase perception," J. Acoust. Soc. Am. **82**, 1560–1586.

Patterson, R. D. (**1994a**). "The sound of a sinusoid: Spectral models," J. Acoust. Soc. Am. **96**, 1409–1418.

Patterson, R. D. (**1994b**). "The sound of a sinusoid: Time-interval models," J. Acoust. Soc. Am. **96**, 1419–1428.

Patterson, R. D., and Akeroyd, M. A. (**1995**). "Time-interval patterns and sound quality," in *Advances in Hearing Research: Proceedings of the 10th International Symposium on Hearing*, edited by G. Manley, G. Klump, C. Koppl, H. Fastl, and H. Oeckinghaus (World Scientific, Singapore, in press).

Patterson, R. D., Anderson, T., and Allerhand, M. (**1994**). "The auditory image model as a preprocessor for spoken language," in Proceedings of the Third ICSLP, Yokohama, Japan, pp. 1395–1398.

Patterson, R. D., Milroy, R., and Allerhand, M. (**1993**). "What is the octave of a harmonically rich note?," in *Proceedings of the 2nd International Conference on Music and the Cognitive Sciences*, edited by I. Cross and I. Deliege (Harwood, Switzerland), pp. 69–81.

Patterson, R. D., and Moore, B. C. J. (**1986**). "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, edited by B. C. J. Moore (Academic, London), pp. 123–177.

Patterson, R. D., Holdsworth, J., and Allerhand M. (**1992a**). "Auditory models as preprocessors for speech recognition," in *The Auditory Processing of Speech: From the Auditory Periphery to Words*, edited by M. E. H. Schouten (Mouton de Gruyter, Berlin), pp. 67–83.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand M. (**1992b**). "Complex sounds and auditory images," in *Auditory Physiology and Perception*, edited by Y. Cazals, L. Demany, and K. Horner (Pergamon, Oxford), pp. 429–446.

Slaney, M., and Lyon, R. F. (**1990**). "A perceptual pitch detector," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, New Mexico, April 1990.