Neural system identification model of human sound localization

Craig Jin
Auditory Neuroscience Laboratory, Department of Physiology, and Computer and Engineering Laboratory, Department of Electrical and Information Engineering, The University of Sydney, New South Wales 2006, Australia

Markus Schenkel
Computer Engineering Laboratory, Department of Electrical and Information Engineering, The University of Sydney, New South Wales 2006, Australia

Simon Carlile
Auditory Neuroscience Laboratory, Department of Physiology, and Institute for Biomedical Research, The University of Sydney, New South Wales 2006, Australia

(Received 19 May 1999; revised 23 February 2000; accepted 9 June 2000)

This paper examines the role of biological constraints in the human auditory localization process. A psychophysical and neural system modeling approach was undertaken in which performance comparisons between competing models and a human subject explore the relevant biologically plausible “realism constraints.” The directional acoustical cues, upon which sound localization is based, were derived from the human subject’s head-related transfer functions (HRTFs). Sound stimuli were generated by convolving bandpass noise with the HRTFs and were presented to both the subject and the model. The input stimuli to the model were processed using the Auditory Image Model of cochlear processing. The cochlear data were then analyzed by a time-delay neural network which integrated temporal and spectral information to determine the spatial location of the sound source. The combined cochlear model and neural network provided a system model of the sound localization process. Aspects of humanlike localization performance were qualitatively achieved for broadband and bandpass stimuli when the model architecture incorporated frequency division (i.e., the progressive integration of information across the different frequency channels) and was trained using variable bandwidth and center-frequency sounds. Results indicate that both issues are relevant to human sound localization performance. © 2000 Acoustical Society of America.

PACS numbers: 43.66.Qp, 43.66.Ba, 43.66.Pn [DWG]

I. INTRODUCTION

A. Human sound localization

The ability to accurately estimate the position of a sound source based upon the available acoustical cues has obvious evolutionary advantages in terms of avoiding predators and finding prey. Indeed, humans are amazingly accurate in their ability to localize broadband sounds. Cues to a sound’s location include the interaural differences in time of arrival (ITDs), the interaural differences in sound level (ILDs), and the monaural spectral cues (recent reviews: Middlebrookes and Green, 1991; Wightman and Kistler, 1993; King and Carlile, 1995). There has been a considerable amount of psychoacoustical research into the auditory processes involved in human sound localization (for a recent review with emphasis on virtual space acoustics see Carlile, 1996a). Furthermore, numerous models of the human sound localization process, as well as that for other animals, have been proposed [Searle et al., 1976; Middlebrooks, 1992; Neti et al., 1992a (for the cat); Backman and Karjalainen, 1993; Nandy et al., 1993; Rosen et al., 1994 (for the barn owl); Chau and Duda, 1995; Datum et al., 1996; Duda, 1997; Janko et al., 1997; Macpherson, 1997; Hofman and Opstal, 1998]. However, there still remains a large gap between the psychophysical and the model explanations (but see, Macpherson, 1997; Janko et al., 1997; Hofman and Van Opstal, 1998). Principal congruence between the two approaches exists for localization performance under restricted conditions, such as for narrow-band sounds where spectral integration is not required, or for restricted regions of space (see Blauert, 1997). Unfortunately, there is no existing computational model that accounts well for human sound localization performance for a wide range of sounds (e.g., varying in bandwidth and center frequency). Furthermore, the biological constraints pertinent to the sound localization process have generally not been explored by these models. Such constraints include the spectral resolution of the auditory system in terms of the number of critical band filters or frequency channels and the role of tonotopic processing and frequency division (progressive integration of information across frequency channels). In addition, the performance requirements of such a system are substantial and involve, for example, the accommodation of spectrally complex sounds, the robustness to irregularity in the sound source spectrum (Macpherson, 1996), and the channel-based structure of spatial coding as evidenced by auditory spatial aftereffects (Hyams and Carlile, 1996). The
The crux of the matter is the notion that "meaningful realism," if built into a model, provides for a better understanding of the underlying processes.

This paper attempts to bridge part of this gap between the modeling and psychophysics. It describes the development of a time-delay neural network model that integrates both spectral and temporal cues for auditory sound localization and compares the results with the corresponding human psychophysical evidence. Realistic constraints have been incorporated into the modeling process by imposing frequency division on the network structure and using training sounds with random center frequency and bandwidth. It is the primary purpose of this paper to examine the relevance of these two issues to human sound localization performance. Additionally, the relationship of these models to a matched filtering algorithm is examined, as is the degradation of localization performance with varying sound level.

With respect to the first issue of frequency division, previous studies have generally neglected its role in sound localization and also its relationship with the integration of spectral information. Nonetheless, the auditory system generally integrates information across frequency to derive the location of a sound source (described below) and it still remains an open question as to just how this integration is accomplished. What is meant by frequency division in this work is that the first computational layer of the network model is not fully connected to the input cochlear frequency channels. Instead, the network structure is constrained such that a progressive integration of information across the different frequency channels is enforced by the network structure as information flows through the different computational layers from the input to the output. Other localization models have certainly used a tonotopic arrangement for their input, but not for the connections of their network model. Typically (e.g., Neti et al., 1992) all of the neurons in the first computational layer are fully connected to all of the input cochlear frequency channels.

The second issue takes a look at the challenging problem of training a network model to localize sounds with varying bandwidth and center frequency. The issue touches upon the question of how frequency content influences the integration of spectral information for sound localization. Furthermore, results suggest that networks with frequency division may have a performance advantage in coping with varying sound conditions.

The model described here is primarily a "black-box" or system identification model. Some of the general advantages of the system identification approach using neural networks are well described by David Zipser (1992), who first introduced the concept of "realism constraints" into the process of neural system identification. As he states (p. 861), "The neural system identification paradigm provides a systematic way to generate realistic models starting with a high-level description of a hypothesized computation and some architectural and physiological constraints." System identification is concerned with replicating the input-output behavior of the system being modeled and neural system identification attempts to model the nervous system in a similar fashion using supervised artificial neural networks. Figure 1 shows a simple diagram illustrating the neural system identification paradigm as it has been applied to the model discussed in this paper. A number of different network architectures and structures for the training data were examined and the results were compared with empirical localization data before making further adjustments to the model. Furthermore, the "black-box" approach was also used to indirectly address two complicated issues related to the sound localization problem: (i) identifying spectral features related to source location and (ii) developing a formal encoding of spatial location (described below).

Modeling the head as a simple sphere, it has been shown that two prominent cues to a sound's location are the interaural time and level cues. The inherent spatial ambiguity derived from such interaural time and level cues has become known as the "cone of confusion" (Wallach, 1939; Blauert, 1997). A further set of cues is provided by the pinna and concha which preferentially amplify or attenuate different frequency components of the sound spectrum depending on the spatial location of the sound source. Psychoacoustical work has demonstrated the importance of these spectral cues

---

**FIG. 1.** The neural system modeling paradigm.
in resolving the ambiguity of the interaural time and level cues (Batteau, 1967; Musicant and Butler, 1984; Oldfield and Parker, 1984), in localization within the vertical dimension (Roffr and Butler, 1968; Gardner and Gardner, 1973; Searle et al., 1975) and in generating the percept of externalized sounds (Durlach et al., 1992). It is generally accepted that the relevant acoustical cues to a sound’s location are described by the head-related transfer function (HRTF) which is typically represented by a finite-length impulse response (FIR) filter (Wightman and Kistler, 1989a; Möller, 1992; Pralong and Carlile, 1994) measured in an anechoic space (recent review: Pralong and Carlile, 1996). Sounds electronically filtered with the HRTF should be localizable when played over earphones which bypass the acoustical filtering of the outer ear. This has been demonstrated (Wightman and Kistler, 1989b; Pralong and Carlile, 1996) and the illusion of free-field sounds using headphones is known as virtual auditory space.

B. Outer ear filtering and spatial location

HRTFs allow for model explorations of localization processing that combine spectral cues along with the interaural time and level cues (Middlebrooks, 1992; Macpherson, 1994). As HRTFs are highly individualized (Wenzel et al., 1988), the modeling results can be compared with human localization performance on an individual basis so as to extract general principles of localization processing via the specifics of individual behavior. Thus a major criterion differentiating more recent sound localization models is their ability to incorporate the individualized spectral filtering of the outer ear. Lyon (1983) and Lindemann (1986) provide early models of the localization process that neglect outer ear filtering but do incorporate interaural time and level cues. Their models derive from the Jeffress (1948) type of interaural cross-correlation model (recent review: Kuwada et al., 1997). Both the Lyon and Lindemann models emphasize the connection between sound localization and sound separation and rely upon temporal correlation applied to the signals at the two ears. Gaik (1993) amended Lindemann’s model to incorporate differences in level across frequency channels by adding weighting elements derived from HRTFs. Unfortunately, Gaik’s model remains constrained to lateralization; no estimate of source elevation is made. Experimentation with the small desktop Koala robots (van Schaik, 1998; Shamma et al., 1998) has highlighted the difficulties associated with navigating toward a sound source in the absence of an estimate of source elevation. Basically it was found that as the source was moved off the horizontal plane, the robots underestimated the lateral position of the source and moved in several iterative and halting steps.

As HRTF recordings became more common, models of the sound localization process began to estimate source elevation by incorporating information relating to the spectral filtering of the outer ear. These models can mainly be classified into two categories. One class of algorithms uses a mathematically well-defined distance measure to directly match the input sound with a database of HRTFs that are associated with specific directions in space (Middlebrooks, 1992; Nandy et al., 1993; Chau and Duda, 1995; Duda, 1997; Macpherson, 1997; Hofman and Opstal, 1998). Another class of algorithms employs an optimization process (often using neural networks) to determine the matching between an input sound and a given location in space (Neti et al., 1992; Backman and Karjalainen, 1993; Rosen et al., 1994; Datum et al., 1996; Janko et al., 1997). These algorithms are similar to each other and provide for an encoding of space based on outer ear filtering.

The difficulties associated with incorporating spectral filtering into the localization process occur at the second or algorithmic level of David Marr’s (1982) hierarchy of explanation for perceptual mechanisms. These difficulties arise because the feature detection paradigm, despite its striking and overwhelming historical success in accounting for sensory phenomena, is difficult to apply when the computational quantities and algorithm are not easily separated. In other words, identifying the exact spectral features used in sound localization is inherently difficult. Generally it is assumed that the most salient spectral cues would be used by the auditory system. Zakarauskas and Cynader (1993) have directly addressed this issue. They propose two possible spectral features that are mathematically specified by first and second order zero-mean difference operators as applied to the spectral signal (see also Carlile, 1990). More recently, Macpherson (1997) has investigated the use of the gradient of the spectral signal. Zakarauskas and Cynader (1993) suggested zero-mean difference operations because they remain invariant to level manipulations. Other methods, which also maintain level invariance, either identify local features in the signal or apply some sort of normalization to the signal.

Possible spectral features that have been suggested as useful for sound localization are spectral notches (Bloom, 1977; May and Huang, 1997) and covert peaks (Musicant and Butler, 1984). For the cat, there is evidence that notch detecting neurons have been found in the dorsal cochlear nucleus (Young et al., 1992) and neural modeling using the cat’s external ear transfer functions by Neti et al. (1992) revealed that some neurons in the network’s hidden layer behave similarly to such notch detecting neurons. Unfortunately the situation may not be so simple for humans. Recent psychoacoustical work by Macpherson (1994, 1997) using irregularly shaped broadband sounds indicates that local features such as notches and peaks are too simple to account for human sound localization behavior. The determination of a specific algorithm to process spectral features was overstepped in this work by employing a time-delay neural network. The neural network optimizes internal network weights that correspond indirectly to “spectral features.” Unlike previous network models, this network incorporates both spectral and temporal information within its architecture.

The issue of the neural encoding for auditory space similarly blurs the boundary between a computational quantity and its corresponding algorithm. In other words, the contribution of the different acoustical cues toward the neural encoding of space depends intimately upon the nature of the spatial encoding itself. It is common to assume a description of space using either an azimuth-elevation angle coordinate system or a lateral-polar angle coordinate system. While
such coordinate systems seem natural, they have the subtle property of being discontinuous. In this work, a direct representation of space in terms of a particular coordinate system was avoided by coding for location directly. That is to say, a local encoding of space was used in which neurons representing locations nearer to the sound source have a higher probability of firing (see Sec. II E 4). This derives from the well-established paradigm that the nervous system uses overlapping receptive fields to encode properties of the physical world. Therefore the time-delay neural network was trained to map spectral and temporal “features” of its input neural activity pattern to different locations in space using an encoding that peaked at the output neuron representing the target location of the input sound and then decayed away as one moves to neurons representing more distant locations. Significantly, the only restrictive assumptions placed upon the model by this realism constraint of overlapping receptive fields involve local properties of the encoding of space and not a specific global coordinate mapping.

II. METHODS

A. Overview

The sound localization performance of a human subject was tested in the free-field. The subject was one of the authors, a male of approximately 30 years of age with normal hearing in both ears, as determined by an audiometric examination, and is referred to as subject X. The sound stimuli consisted of three different bandpassed sounds: (1) a low-passed sound with frequencies 300–2000 Hz; (2) a high-passed sound with frequencies 2000–14000 Hz; and (3) a broadband sound with frequencies 300–14000 Hz. These frequency bands respectively cover conditions in which either temporal cues, spectral cues, or both dominate the localization process (the evidence for this fact derives from the “intensity-theory” and “phase-theory” of spatial coding, related to the duplex theory of sound localization; see Carlile, 1996b; Carlile et al., 1999). Such sounds should hopefully allow for the integration of both spectral and temporal information during localization and provide for a rigorous examination of subsequent modeling. The subject performed five localization trials for each sound condition; 76 test locations, evenly distributed on the sphere, were used during each trial. As the responses to the low-pass stimuli showed an unbalanced attraction to the frontal hemisphere, the test conditions were arranged such that the low-pass and broadband stimuli were randomly mixed in equal proportions, while the trials with high-pass stimuli were performed in straight sets. Therefore in order to obtain a single trial result for the low-pass or broadband conditions, the subject needed to perform two tests with randomly mixed sounds. The trial results for an individual condition were then obtained by unscrambling the two tests.

Numerous studies have been conducted in which subjects are presented with spectrally restricted sound stimuli, as above, in an effort to clarify the role of the spectral cues in the localization process (e.g., Carlile et al., 1999; Blauert, 1997; Butler, 1986; King and Oldfield, 1997; Middlebrooks and Green, 1992). These studies have shown that accurate sound localization requires a relatively large signal bandwidth (King and Oldfield, 1997) and that subjects tend to make systematic localization errors on spectrally restricted sounds (Carlile et al., 1999). Furthermore, it is generally believed that the localization errors are related to the spectral filtering characteristics of the auditory periphery (Blauert, 1997; Middlebrooks and Green, 1992). Using sound stimuli similar to those described above, Carlile et al. (1999) have demonstrated that the mislocalizations of different subjects varied, but that for a given subject these localization errors were systematic. It is thus a basic premise of this work that the subject’s systematic mislocalizations (demonstrated below) should be related in a meaningful way to the acoustical properties of his auditory periphery.

In order to incorporate the acoustical filtering of the outer ear into the modeling process, measurements were made of the subject’s HRTFs at approximately 400 directions evenly distributed on the sphere. Directional transfer functions (DTFs) were then calculated from the HRTFs as in Middlebrooks and Green (1990). These DTFs were then used in the modeling process to simulate the directional aspects of the acoustical filtering of the outer ear. To establish that the DTFs appropriately indicate the direction of a sound source and were therefore an accurate input to the model, the subject repeated the identical localization task as above, but in a virtual sound environment. In this sound environment, the sound stimuli were filtered with the DTFs and played over in-ear tube phones so that the subject’s cues to the location of the virtual sound source came entirely from the DTFs.

The sound localization process was then modeled using two basic system components: (1) a modified version of the physiological Auditory Image Model (Patterson and Allerhand, 1995; Giguère and Woodward, 1994) which simulates the spectro-temporal characteristics of peripheral auditory processing and (2) the computational architecture of a time-delay neural network (TDNN). A TDNN was chosen because of its ability to process and weight both spectral and temporal cues in a biologically plausible manner. Previous work has used neural networks without time delays, but with a number of inputs devoted to an ITD cross-correlation cue (e.g., see Backman and Karjalainen, 1993; Chung et al., 2000) and other inputs devoted to the separate frequency channels corresponding to the spectral cues. Another technique, used by Macpherson (1997), is to artificially but equally weight the ITD cue versus the spectral cues.

B. Measuring sound localization performance

A detailed presentation of the methods used to train and test the subject in free-field sound localization as well as generate and present sound stimuli can be found in Carlile et al. (1997a). A short summary of the basic techniques is presented here.

1. The testing environment

The human localization experiments were carried out in a darkened anechoic chamber. Free-field sound stimuli were presented from a loudspeaker carried on a semicircular robotic arm. This arrangement allowed for placement of the speaker at almost any location on the surface of an imaginary
sphere, one meter in radius, centered on the subject’s head. The perceived location of the sound source was indicated by the subject turning squarely to face the sound and then pointing his nose in the direction of the perceived source. The subject’s head orientation and position were monitored using an electromagnetic sensor system (Polhemus, Inc.) consisting of a transmitter and receiver, the latter fastened to a headband worn by the subject. Between localization trials, the subject was required to align himself with the calibrated start position (azimuth 0°, elevation 0°) so that each stimulus was oriented correctly. This task was aided by a grid of colored light emitting diodes indicating instantaneous alignment error. A hand-held response button was used by the subject to both initiate a trial and indicate his completion of the localization task.

2. Stimulus generation

The noise stimuli were generated using D/A conversion at 80 kHz and delivered to a power amplifier (Quad 306) via a programmable attenuator (TDT: PA4). The loudspeaker (VIFA No. D267TG-35), mounted on the hoop-positioning system, had a response characteristic that was spectrally flat within 10 dB between 2 and 16 kHz and spectrally flat within 10 dB. Between 2 kHz and 300 Hz, the speaker had a high-pass roll-off of approximately 8 dB per octave. The sound stimuli were presented at approximately 70 dB SPL measured at the center of the positioning system. A “fresh” white Gaussian noise was appropriately bandpassed for each trial. The duration of the stimulus was restricted to 150 ms (with 10-ms raised-cosine onset and offset ramps) to ensure that head movement was not initiated during the presentation (Thurlow and Mergener, 1970; Carlile et al., 1997).

C. Measurement of the directional transfer functions

Measurements of the subject’s outer ear filter functions were carried out in the anechoic chamber. The measurements were made for both ears simultaneously using a “blocked ear” technique with a small Knowles electret microphone (EA–1954) placed in each ear (Möller et al., 1995). Three hundred ninety-three measurements were made at locations evenly distributed on the sphere. In order to improve the signal-to-noise ratio, digitally constructed Golay codes with a 1024 impulse length were used as the recording stimulus (see Zhou et al., 1992). The response of the microphone was bandpass filtered from 200 Hz to 16 kHz, digitized at 80 kHz, and averaged over 16 repetitions of the stimulus.

The transfer function of the outer ear was obtained by deconvolving the response of the microphone in the free-field from the response recorded at the entrance to the blocked ear (see Mehdgardt and Mellert, 1977; Wightman and Kistler, 1989b; Pralong and Carlile, 1994). The free-field calibration of the probe microphones was recorded without the subject in the anechoic chamber and at the position corresponding to the center of the subject’s head. Following the deconvolution, the directional transfer functions (DTFs) were calculated as in Middlebrooks and Green (1990).

D. Evaluation of the directional transfer functions

A detailed presentation of the methods used to evaluate a subject’s measured DTF recordings can be found in Pralong et al. (1996a). The DTFs should accurately capture, for a particular individual, all of the relevant acoustical cues to the direction of a sound source. Therefore sounds which are electronically filtered with the appropriate DTF for a particular direction in space should appear to come from that direction in space when played over headphones. To assess the quality of the DTF measurements, the subject repeated the previous localization test in virtual auditory space. The virtual sound stimuli were constructed by convolving the measured DTFs with the original sound source and then playing this sound to the subject using tubephones (ER-2, Etymotic Research). The tubephones are designed to produce an approximately flat frequency response, within 3 dB, at the human ear drum over the frequency range 200 Hz to 16 kHz.

E. A network model of sound localization

1. Overview of the modeling process

The sound localization model consisted of two basic system components: (1) a modified version of the physiological Auditory Image Model which simulates the spectrotemporal characteristics of peripheral auditory processing, and (2) the computational architecture of a time-delay neural network. The modeling process can be broken down into four stages as shown in Fig. 2. In the first stage a sound stimulus was generated with specific bandpass characteristics. The sound stimulus was then filtered with the subject’s right and left ear DTFs to render an auditory stimulus originating from a particular location in space. The psychophysical data indicate that this stimulus is a good approximation to the real sound that would be heard by the subject in the free-field (see Fig. 7 below; Zahorik et al., 1996; Hartmann and Wittenberg, 1996). The auditory stimulus was then processed by the Auditory Image Model (AIM) to generate a neural activity profile that simulates the output of the inner hair cells in the organ of Corti. This neural activity profile indicates the spiking probability of auditory nerve fibers. Finally, in the fourth and last stage, a time-delay neural network computed the spatial direction of the sound input based on the distribution of neural activity calculated by AIM.

2. Peripheral auditory processing of the sound input

The sound stimuli presented to the model were filtered with the subject’s DTFs in the same manner as that used for evaluating the subject’s sound localization ability in a virtual sound space (see Sec. II.D). Following filtering, the physiological version of AIM, as opposed to the functional version, was used to process the stimulus because of its greater biological plausibility and its ability to include sound level dependencies. The standard physiological version includes its own external ear filtering. Therefore, with the aid of Christian Giguère (an author of the original software code), a
software switch was implemented to disable this filtering under these conditions. The AIM model simulated the following: (1) the transmission through the middle ear; (2) the basilar membrane motion and cochlear hydrodynamics; (3) the fast motile mechanism of the outer hair cells; and (4) the neural transduction process of the inner hair cells. The distribution of cochlear filters across frequency was chosen such that the minimum center frequency was 300 Hz and the maximum center frequency was 14 kHz, with the density of filters being 5 filters per critical band, giving a total of 153 filters essentially equally spaced on a logarithmic scale (roughly 5 or more filters per critical band is required by the physiological version of AIM). A sampling frequency of 80 kHz was chosen to match the sampling frequency of the subject’s HRTFs.

The output of the AIM model was downsampled from 80 to 6.7 kHz and the number of output channels reduced from 153 to 31, resulting in approximately one channel per critical band. Thus the auditory preprocessing was performed with high spectral and temporal resolution and reduction occurred only at the level of the neural activity profile. This downsampling amounts to making implicit assumptions (“realism constraints”) about the temporal and spectral resolution of the underlying physiology. On the assumption that the auditory nerve contains phase information up to 3 kHz, a 6.7-kHz sample rate should be suitable for capturing ongoing ITD information. Higher temporal resolution, however, would be required to effectively render the onset time information. As for the reduction in spectral resolution, the 31 cochlear channels seemed sufficient to produce reasonable localization performance with a manageable amount of computation. This spectral resolution is compared with that of other published models in Table I. To our knowledge, no systematic modeling has been performed to justify any particular density of frequency channels, although psychophysical evidence seems to indicate that human sound localization is surprisingly robust to reductions in spectral resolution (Leung et al., 1998; Kulkarni and Colburn, 1998).

### 3. The time-delay neural network

As a time-delay neural network was used to model the auditory system’s sound localization process, a brief review of TDNNs is given (for a more complete description see Schenkel, 1995; Wan, 1993). In traditional artificial neural network models, the summing nodes of the neurons correspond to a simple static weighted sum and the connections via the summing nodes provide a spatial representation of the signal flow through the network. A fundamental aspect of the TDNN is that it essentially adds temporal filtering to the summing nodes of the neurons, thus incorporating a temporal dimension into the network computation. The temporal filtering is implemented as a discrete finite impulse response (FIR) filter,

\[
s(k) = \sum_{n=0}^{N} w(n)x(k-n),
\]

where \( k \) is a discrete time index, \( N \) is the length of the FIR filter, \( x(k) \) is the input, and \( s(k) \) is the filter output. This operation is shown in Fig. 3 using the standard representation of a tapped delay-line for the FIR filter. The output \( y(k) \) from the node is then calculated by performing a soft thresholding operation using, in this case, the hyperbolic tangent,

\[
y(k) = \tanh(s(k)).
\]

Each connection node in the TDNN is formed using a FIR filter, such as described above. Therefore, a TDNN can be described as a multi-layer feed-forward neural network in which the outputs of a layer are buffered in an internal memory for several time steps and then connected to the next layer (see Fig. 4). The training of a TDNN involves adapting the FIR filter coefficients at the connection nodes to mini-

---

**TABLE I. Comparison of different models and their spectral resolution.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Begin frequency (kHz)</th>
<th>End frequency (kHz)</th>
<th>Frequency channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neti et al., 1992 (binaural)</td>
<td>2</td>
<td>14</td>
<td>44</td>
</tr>
<tr>
<td>Neti et al., 1992 (monaural)</td>
<td>2</td>
<td>14</td>
<td>88</td>
</tr>
<tr>
<td>Middlebrooks, 1992</td>
<td>3</td>
<td>16</td>
<td>40</td>
</tr>
<tr>
<td>Zakarauskas et al., 1993</td>
<td>1</td>
<td>22</td>
<td>104</td>
</tr>
<tr>
<td>Chau and Duda, 1995</td>
<td>2</td>
<td>22</td>
<td>40</td>
</tr>
<tr>
<td>Janko et al., 1997</td>
<td>0.38</td>
<td>17</td>
<td>23</td>
</tr>
<tr>
<td>Current model</td>
<td>0.3</td>
<td>14</td>
<td>31</td>
</tr>
</tbody>
</table>
mize the computational error. An algorithm known as temporal backpropagation can be used to train the TDNN. The temporal backpropagation algorithm is akin to the traditional backpropagation algorithm and one formulation of the algorithm (see Wan, 1993) describes the backpropagated delta terms (as is commonly described in the regular backpropagation algorithm) as being calculated by filtering the error “backwards” through the FIR filter, with the reverse FIR filter described as shown in Fig. 3(b).

The layers of a TDNN, as described above, perform successively higher-level feature extraction. TDNNs have previously been applied to speech recognition and are well suited to sequential signal processing tasks (Waibel et al., 1989; Lang and Hinton, 1988). The input fields of the neurons are restricted to a time window of constant duration and generally process the input signal sequentially, e.g., the network “moves across” the input. There are thus far fewer weights in this architecture than there would be in a fully connected network that processed the full extent of the input signal. In this case, the input to the TDNN was the neural activity pattern produced by the Auditory Image Model, consisting of the temporal signals in the different cochlear channels for the left ear and similarly for the right ear. Any given input neuron would then detect a particular local feature in the input signal (essentially the spiking probability of the different auditory nerve fibers). The receptive field of the neurons in the first computational layer was typically restricted to a time window of 2.25 ms. This time window is of the same order of magnitude as the spectral integration time of 5 ms measured psychophysically using FM sweeps by Hofman and Van Opstal (1998). As the neurons move across the time axis, they detect the presence or absence of a particular local feature. By using several neurons at each time-step, the network detects numerous different features. This operation corresponds in essence to a convolution of the weight kernels of the neurons with the input. The time component of the input representation is gradually eliminated by subsampling the convolution at each layer. To partially compensate for the loss of information, the network usually and gradually increases the number of features. Such a structure is referred to as a bi-pyramidal network architecture and progressively converts time information into feature information. Through this structure, the activity of neurons at progressively higher levels depends on increasingly larger parts of the input. In this work, the last layer of the TDNN consisted of as many neurons as there were different classes of spatial directions. Thus, output scores were produced for each of 393 directions in space, evenly distributed on a sphere, with approximately 10 degrees of angular spread between directions. These scores can be interpreted as the relative probabilities that the given sound came from that particular direction in space.

4. A neural encoding of space

The spatial location of a sound source was encoded by the network as a distribution of neural activity in its output layer. Following Neti et al. (1992), this encoding was modeled on empirical data obtained from the superior colliculus of a cat, as demonstrated by Middlebrooks and Knudsen (1984). Such an encoding entails a distributed response with the peak occurring at the output neuron representing the target location of the input sound. The output response would then decay away in the form of a two-dimensional Gaussian as one moves to neurons farther away from the target location. It has been implicitly assumed that the output response has isotropic resolution on the sphere. The neural activity of a given output neuron was calculated, during training of the network, according to the following formula:

\[ \gamma_i = e^{(\ln(0.9) \cdot (l_i^2 + \kappa^2))}, \]

where \( \gamma_i \) is the output activity for the \( i \)th neuron, \( l_i \) is the spherical angle between the sound input location and the location represented by the \( i \)th neuron and \( \kappa \) is a constant spherical angle determining the inherent precision in the output. As given by the formula, the output neural activity is scaled between 0 and 1 and will be within 90% of full value.
III. NETWORK ARCHITECTURES

The major auditory brainstem nuclei demonstrate substantial frequency division within their structure (see Irvine, 1992). In other words, the systematic frequency organization of the primary auditory nerve fibers that innervate the cochlea seems to carry forward to the nuclei within the brainstem’s auditory pathway. This arrangement is described as a tonotopic organization. Despite this fact and to our knowledge, no previous network model for sound localization incorporates such frequency division within its architecture. Therefore, in this work, three different architectures were examined with varying amounts of frequency division imposed upon the network structure. The different architectures are described below. Importantly, a network architecture was considered different only if there exists a fundamental difference in its structure. So, for example, two networks with a slightly different number of neurons in a hidden layer were not considered different. 3

A. MLP architecture

As a basis for developing a TDNN architecture, a simple multi-layer perceptron (MLP) architecture without time delays was examined. The input to the MLP was calculated as the temporal average of the input data for the TDNNs. In other words, the output of each channel of the AIM is averaged over the duration of the stimulus so that the MLP gets one value per channel per ear. The number of neurons in each layer of the MLP is given in Table II and each layer of the network was fully connected to the previous layer (i.e., there was no frequency division within its structure).

B. TDNN architecture A

A TDNN architecture was created with a similar structure to the MLP, such that the relative number of neurons in different layers of the TDNN network were the same as that for the MLP. The description of a TDNN architecture is more complex than that for an MLP. In order to fully describe a computational layer of the TDNN, four characteristic numbers must be specified: (1) the number of neurons; (2) the kernel length, a number which determines the size of the current layer’s time-window in terms of the number of time steps of the previous layer; (3) the kernel width, a number which specifies how many neurons in the previous layer with which there are actual connections; and (4) the undersampling factor, a number describing the multiplicative factor by which the current layer’s time-step interval is increased from the previous layer’s. Using this nomenclature, the architecture of the first TDNN, henceforth referred to as architecture A, is summarized in Table III and each layer of the network was fully connected to the previous layer.

C. TDNN architecture B

The brainstem nuclei involved in processing ILDs and ITDs are physiologically distinct with the separate streams represented in the lateral superior olive and the medial superior olive (Irvine, 1992). Therefore it was decided to impose a similar network structure on the second TDNN architecture, architecture B. This TDNN essentially consisted of the combination of two temporary and separate networks that were trained individually on low-pass (300 Hz to 2 kHz) and high-pass (2 kHz to 14 kHz) sounds. The assumption was that the network trained on low-pass sounds would emphasize the ITD information contained within the low-frequency cochlear channels, while the network trained on high-pass sounds would emphasize the ILD information contained within the high-frequency cochlear channels. The first hidden layers of both these networks were then combined to produce a hidden layer with twice as many neurons, which was then used as the first hidden layer of the TDNN with

![Figure 5: Target neural activity for a sound located at zero degrees of azimuth and elevation.](Image)

**TABLE II. The MLP network.**

<table>
<thead>
<tr>
<th>Layer</th>
<th>Neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>62</td>
</tr>
<tr>
<td>Hidden 1</td>
<td>40</td>
</tr>
<tr>
<td>Hidden 2</td>
<td>20</td>
</tr>
<tr>
<td>Output</td>
<td>393</td>
</tr>
</tbody>
</table>

**TABLE III. The TDNN with architecture A.**

<table>
<thead>
<tr>
<th>Layer</th>
<th>Neurons</th>
<th>Kernel length</th>
<th>Kernel width</th>
<th>Undersampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>62</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Hidden 1</td>
<td>40</td>
<td>15</td>
<td>62</td>
<td>2</td>
</tr>
<tr>
<td>Hidden 2</td>
<td>20</td>
<td>10</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>Output</td>
<td>393</td>
<td>6</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>
architecture B. Finally, this TDNN was trained identically to the TDNN with architecture A, except that the weights of the first hidden layer (which were trained previously) were held fixed, i.e., they were not updated during training. The detailed structure of the architecture is described in Table IV with the 60 neurons in the first hidden layer resulting from the combination of two layers of 30 neurons. As for architecture A, each layer of the network was fully connected to the previous layer.

**D. TDNN architecture C**

A third TDNN network architecture, architecture C, was developed with more substantial frequency division within its structure than the previous networks. This network’s connections were constrained by frequency in a tonotopic-like arrangement (see Fig. 6). The 31 input cochlear frequency channels were split into ten overlapping groups consisting generally of six contiguous frequency channels. So, for example, the first group contained the frequency channels 1–6, while the second group contained the frequency channels 4–9 and so forth. Each input group of the first hidden layer contained 5 neurons. The kernel widths of these neurons were set, not to the total number of frequency channels in the input layer, but only to the six contiguous frequency channels defining the group. Information across the different groups of frequency channels was progressively integrated in the higher layers of the network. Thus, for example, the second hidden layer contained six groups of neurons connected to four or more different neuronal groups in the first layer. The size of the neuronal groups in the second layer was adjusted so that those spanning a greater frequency range contained more neurons. The exact connection arrangement can be seen in Fig. 6 and the architecture of the different layers is summarized in Table V.

### TABLE IV. The TDNN with architecture B.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Neurons</th>
<th>Kernel length</th>
<th>Kernel width</th>
<th>Undersampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>62</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Hidden 1</td>
<td>60</td>
<td>15</td>
<td>62</td>
<td>2</td>
</tr>
<tr>
<td>Hidden 2</td>
<td>20</td>
<td>15</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>Output</td>
<td>393</td>
<td>3</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

### TABLE V. The TDNN with architecture C.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Neurons</th>
<th>Kernel length</th>
<th>Kernel width</th>
<th>Undersampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>62</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Hidden 1</td>
<td>50</td>
<td>15</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Hidden 2</td>
<td>28</td>
<td>10</td>
<td>4,5,6</td>
<td>2</td>
</tr>
<tr>
<td>Output</td>
<td>393</td>
<td>4</td>
<td>28</td>
<td>1</td>
</tr>
</tbody>
</table>

**IV. TRAINING NETWORKS TO LOCALIZE SOUNDS**

**A. Learning algorithm**

In terms of a neural system modeling paradigm, supervised training of a neural network is an optimization procedure, which to a first approximation should not depend on the learning algorithm. In other words, if a highly restrictive and specialized algorithm is required to achieve convergence upon a reasonable solution, then the model is likely to be poor and uninformative. In this work, the error backpropagation algorithm was used with a summed squared error measure (the most well-known derivation is given by Rumelhart et al., 1986). Typically a training cycle was initiated with a learning rate of 0.05 and was decremented in stages with the training continued until a plateau was reached in the learning curve. The final average error per output neuron for a network solution depended on the structure of the training data with a typical value being 0.1 per neuron.

**B. Structure of the training data**

More relevant perhaps than the training algorithm, from a modeling point of view, is the learning task itself and the actual structure of the training data. In this work, one objective was to better approximate listening conditions in which a multitude of sounds are possible. Such conditions are, from an ecological point of view, arguably more plausible than listening conditions in which, for example, only broadband sounds are presented. In an effort to better approximate such listening conditions, a training method that uses sounds with varying bandwidth and center frequency was used. Such a training method presents a challenging problem because the network must learn to localize a sound based on features that are not certain, but only probable given the sound environment. For example, the large spectral contrast at the band-edge of a bandpass sound is not truly indicative of the sound’s location. However, if a spectral notch in an HRTF

![Figure 6. Tonotopic arrangement of neurons for architecture C.](image-url)
for a given location occurs at the frequency of the band-edge, then large spectral contrast at that frequency can sometimes be indicative of the sound’s location. The notion of “probable” spectral features is well described by Rogers and Butler (1992) and Butler and Musicant (1993).

In order to better approximate listening conditions with varying sound stimuli, two different sound arrangements were used for training the networks. The first arrangement consisted of three types of stimuli embodying the test conditions: low-pass, high-pass, and broadband noise. These were presented with equal one-third probability. In a second arrangement, the center frequency and bandwidth of the noise were chosen randomly. That is to say, the bandwidth was chosen uniformly and randomly from the number of allowable cochlear channels (i.e., a number between 2 and 31 channels inclusive). The center frequency was then chosen uniformly and randomly from those cochlear channels that could accommodate the given bandwidth. In this way, the band-edge frequencies were always chosen to be within the frequency range 300–14000 Hz. For example, suppose that a bandwidth of 15 cochlear channels was chosen. The center frequency would then be chosen from the cochlear channels between 8 and 24 inclusive. Thus the noise bandwidths were based on the same logarithmic frequency scale as the cochlear channels.

Additionally, one had to ensure that an appropriate number of training patterns was used. A list of the total number of weights and the Vapnik-Chervonenkis (VC) dimension for the different networks is given in Table VI. The VC dimension is a scalar value providing a measure of the intrinsic storage capacity of a network, i.e., the number of random patterns that can be stored and was calculated as described in Baum and Haussler (1989). As shown, the number of connection weights, or free parameters of the model, is large and varies between approximately 10,000 and 100,000. Therefore an even larger number of patterns was required during training; approximately 200,000 patterns were used (see Table VI). The training stimuli generally consisted of 2800 directional sounds and the estimated equivalent number of training patterns, shown in Table VI, was calculated based on the double Gaussian output encoding and the convolutional structure of the TDNNs. In actuality, the most practical criterion for determining an appropriate number of training patterns was to empirically observe the point at which the network’s localization performance on unseen test sounds was equivalent to that for similar training sounds.

C. Frequency selective training

In association with the varying frequency structure of the training data and the tonotopic network architecture, the backpropagation training algorithm automatically provides, in this case, a frequency selective learning algorithm. It is a feature of the learning rule that a neuron’s weights are only updated when there is input activity in the corresponding connections. As the input cochlear frequency channels are connected to separate groups of neurons in the first computational layer, some bandpass sounds will not activate all of the neuronal groups. So, for example, a training sound containing only low frequencies will not train the high-frequency neurons and vice versa.

V. METHODS OF ANALYSIS

A. Quadrature view

The extent of the qualitative differences in the localization performance of the competing models was sufficient to clearly distinguish between them. The distribution of localization responses are presented on four different spherical plots, referred to as a quadrature view. The four plots illustrate the front, back, left, and right hemispheres of space looking in toward the subject. Each perspective shows a hemisphere of space as seen from the outside looking in (e.g., see Fig. 7). Thus the localization response for any sound target located in the frontal hemisphere is shown on the “front” plot. Similarly, the response for any target location toward the rear of the subject is shown on the “back” plot and so on. The target locations are indicated by crosses and, depending upon the number of localization responses for a given target, either the response or the centroid of a number of responses are denoted by a filled circle. A solid line is drawn connecting the target and response locations along a great circle.

B. Spherical statistics

Methods of spherical statistics were used to analyze the experimental psychophysical data. The localization responses of the subject for a given bandpass condition and presentation method (e.g., low-pass, high-pass; free-field, virtual space) were pooled across the five trials for that condition. For a specific target location, the mean direction of the localization judgments was calculated in terms of the “centroid,” a measure of central tendency in spherical statistics (see Fisher et al., 1987). In addition, the standard deviation about the centroid was modeled using either a Fisher or Kent distribution, depending upon whether the data points were symmetrically distributed (Fisher) or best fit by two principal components (Kent). A detailed explanation of these methods is given in Leong and Carlile (1998). Thus when the subject’s localization performance is shown in quadrature view, the standard deviation about the centroid is depicted by an ellipse with the length of the major and minor axes indicating the amount of variation.

VI. RESULTS

A. Localization performance of the subject

The subject’s localization performance in the free-field as compared with that in the virtual auditory space is demonstrated in Fig. 7. Human free-field localization is shown in
the first column, Figs. 7(a)–(c), while the human virtual auditory space localization is shown in the second column, Figs. 7(d)–(f). The three different bandpass conditions are illustrated in three rows and consist of broadband [Figs. 7(a), 7(d)], high-pass [Figs. 7(b), (e)], and low-pass [Figs. 7(c), (f)] noise. As discussed earlier (see Sec. II), the contrast between the subject’s free-field localization and virtual auditory space localization should portray the combined influence of the following differing conditions: (1) The free-field sounds were spectrally flat to within 10 dB between 2 and 16 kHz and there was a rather steep high-pass roll-off (8 dB/octave) between 2 kHz and 300 Hz, while the virtual sound sources were spectrally flat within 3 dB between 300 Hz and 16 kHz. (2) The acoustical cues provided naturally during the free-field localization are embodied in the HRTFs, whereas the acoustical cues provided via electronic filtering during the virtual auditory space localization are embodied in the DTFs. Despite these differing sound conditions, the localization data demonstrate that the performance between the two sound environments was convincingly similar. This suggests that the DTFs were an accurate measure of the psychophysically relevant components of the HRTFs in the free-field condition. The data show that the subject performed well in both the broadband and high-pass sound conditions and rather poorly in the low-pass condition. In both the broadband and high-pass sound conditions [Figs. 7(a), (d); 7(b), (e)], the front-back errors were in the minority of responses and were removed when calculating the standard deviation of the responses about the centroid. Front-back errors were defined as localization responses in which the perceived location crossed the vertical plane through the interaural axis which separates the anterior and posterior hemispheres of space. For the low-pass condition, the subject demonstrated gross mislocalizations, with the responses clustering toward the lower and frontal hemispheres. The distribution of errors for the low-pass condition was again based on the majority response such that front-back differences were removed; the gross mislocalizations correspond mainly to the traditional cone of confusion errors (see Carlile et al., 1999). As a quantitative measure of the similarity between the two sound environments, the spherical correlation coefficient

FIG. 7. The localization responses for subject X in both the free-field and virtual space are plotted in quadrature view showing the front (f), back (b), left (l), and right (r) hemispheres of space. To assist in identifying the spatial orientation, a solid black triangle indicates the nose position. The free-field responses are shown on the left and the virtual auditory space responses are on the right. A cross denotes the target location, a filled circle denotes the centroid of the perceived locations and a solid line joins the two. The distribution of localization errors is shown by an ellipse, whose major and minor axes indicate standard deviation. The spherical correlation coefficient (s.c.c.) is indicated at the bottom of the quadrature plots. (a) Free-field broadband localization. (b) Free-field high-pass localization. (c) Free-field low-pass localization. (d) Virtual source broadband localization. (e) Virtual source high-pass localization. (f) Virtual source low-pass localization.
(see Fisher et al., 1987) between the target and perceived locations was computed for each experiment. These data consisted of five trials with 76 locations, yielding a total of 380 localizations. A coefficient of 0.96 was obtained for the free-field environment with 2% of the responses being front-back errors. In the virtual sound environment, a similarly high correlation coefficient of 0.95 was obtained with 5% of the responses being front-back errors. The high degree of correlation between the target and perceived locations in the virtual sound environment again suggests that the synthesized virtual sound sources were highly effective and, most importantly, are suitable for modeling the localization process. Certainly, however, the increased number of front-back errors that occurred in the virtual sound space suggests that it is not an exact reproduction of the free-field. In this case, it is not known whether these differences were caused by signal differences or behavioral and performance issues. Nonetheless, these quantitative measures indicate that both sound environments conveyed much the same acoustical information with respect to sound location.

B. Localization performance of the MLP

The localization performance of the multi-layer perceptron is shown in Fig. 8. This particular network was trained with equal probability on the three bandpass conditions. The localization responses of the network are plotted side by side with the localization responses of the subject in the free-field. This is meant as an aid to the reader in assessing the model’s performance. The localization data are shown in quadrature view and arranged such that the subject’s responses are in the left column [Figs. 8(a)–(c)] and the model’s responses are on the right [Figs. 8(d)–(f)]. The localization responses for the three bandpass conditions are illustrated in different rows. Only single trial results are shown for the models as a qualitative examination of the accuracy of the localization responses was sufficient, rather than the more detailed and quantitative analysis of the variance of replicate responses. It was for this reason that no effort was made to smoothly interpolate or otherwise refine the output of the network to have a precision better than...
approximately ten degrees of spherical angle, which would have been required for any meaningful interpretation of response variation. The MLP illustrated in Fig. 8 demonstrates good localization performance for both the broadband and high-pass sound conditions and much poorer performance for the low-pass sound condition. The performance of the MLP is quantitatively compared with the human subject in Table VII using the spherical correlation coefficient, the average spherical angle error and the percentage of front-back confusions. There is evident similarity, but with some differences, between the subject’s localization performance and that of the MLP; notably, there are some structural similarities in the low-pass sound mislocalizations. The network’s low-pass sound localization in the left hemisphere of space is sometimes pulled, similarly to that of the subject, toward the lower and frontal hemisphere of space. However, unlike the subject’s localization performance, the MLP demonstrates significant asymmetry between the left and right hemispheres of space for the low-pass stimulus. It is unlikely that the response asymmetry was caused by an error in the model because: (1) the same calculations are performed independently of the stimulus location; (2) the model demonstrates symmetry in its localization performance for the broadband and high-pass sounds. It is also unlikely that there was an error in the sound input because the subject localized the same sounds as the MLP. It can, of course, be argued that the subject’s acoustical filtering at the left and right ear (the HRTFs) is asymmetric, but to a first approximation it is likely that the acoustical filtering is essentially bilaterally symmetric. Thus this asymmetry shown in the low-pass condition may indicate that the training of the network was insufficient to develop meaningful structure in the network weights corresponding to the low-frequency cochlear channels. There is also likely to be some redundancy in the information across the different cochlear frequency channels, which may complicate the sufficient development of the network weights corresponding to the low-frequency cochlear channels. This gives some indication that training methods using sounds with varying frequency content may be an important consideration when examining neural network models of sound localization.

These results, in general, indicate that the methodology used for preprocessing the input sound stimuli was reasonable and that the error backpropagation algorithm could reasonably extract information relating to the source location and thus paved the way for developing the TDNN, a model with a substantially more realistic ability to extract temporal as well as spectral information from the sound stimulus.

### C. Localization performance of the TDNN with architecture A

The localization responses of the TDNN with architecture A, the architecture most similar to the MLP, is shown in Fig. 9. A quantitative comparison of this network’s performance with that of the human subject is also given in Table VII. The training for this network was identical to that for the MLP. Despite the similarities between the two networks, the TDNN failed to localize sounds across the different bandpass conditions. This TDNN shows poor localization performance in the high-pass condition, in contrast to the human subject, and a considerable clustering of responses in the low-pass condition toward the left and right interaural axes, again in contrast to the human subject where the responses cluster at lower frontal elevations. This result is representative of several networks similar in structure to architecture A, but with a variable number of hidden neurons and minor differences in training regimes. This TDNN did, however, show reasonable performance in the broadband sound condition. The poorer localization performance of the TDNN as compared with the MLP could be attributed to the incorporation of temporal information within the frequency channels. As the training methods for the two networks were identical, one must examine the architectures and the input data to find the cause behind the differences. As described in Sec. III, the architectures were the same except for the convolutional structure imposed by the TDNN. The TDNN’s weights are convolved with the input signal over time and thus can incorporate temporal information from the input data into its processing. Given that there is temporal information in the signal, this information could interfere with the processing of the spectral information which results in differences in the localization performance of the two networks. Error in the backpropagation algorithm can be eliminated because of the reasonable performance by the TDNN in localizing broadband sounds. Thus it appears most likely that there has been a failure of the TDNN to robustly incorporate temporal and spectral information into the localization process across the different bandpass conditions. As will be seen in the following sections, this inadequacy to incorporate both temporal and spectral information is removed, to a great extent, by dividing the input frequency channels into different groups and then progressively interconnecting the neurons in the higher layers across frequency.

### TABLE VII. Comparison of the localization performance results for the subject and the competing models.

<table>
<thead>
<tr>
<th>Sound condition</th>
<th>Spherical correlation coefficient</th>
<th>Mean spherical angle error</th>
<th>Percentage front-back error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject X</td>
<td>Broadband</td>
<td>0.92</td>
<td>11</td>
</tr>
<tr>
<td>High-pass</td>
<td></td>
<td>0.90</td>
<td>12</td>
</tr>
<tr>
<td>Low-pass</td>
<td></td>
<td>0.03</td>
<td>50</td>
</tr>
<tr>
<td>MLP</td>
<td>Broadband</td>
<td>0.97</td>
<td>5</td>
</tr>
<tr>
<td>High-pass</td>
<td></td>
<td>0.94</td>
<td>9</td>
</tr>
<tr>
<td>Low-pass</td>
<td></td>
<td>-0.01</td>
<td>53</td>
</tr>
<tr>
<td>TDNN architecture A</td>
<td>Broadband</td>
<td>0.89</td>
<td>9</td>
</tr>
<tr>
<td>High-pass</td>
<td></td>
<td>0.36</td>
<td>35</td>
</tr>
<tr>
<td>Low-pass</td>
<td></td>
<td>0.00</td>
<td>54</td>
</tr>
<tr>
<td>TDNN architecture B</td>
<td>Broadband</td>
<td>0.93</td>
<td>10</td>
</tr>
<tr>
<td>High-pass</td>
<td></td>
<td>0.92</td>
<td>10</td>
</tr>
<tr>
<td>Low-pass</td>
<td></td>
<td>0.25</td>
<td>34</td>
</tr>
<tr>
<td>TDNN architecture C</td>
<td>Broadband</td>
<td>0.94</td>
<td>9</td>
</tr>
<tr>
<td>High-pass</td>
<td></td>
<td>0.92</td>
<td>10</td>
</tr>
<tr>
<td>Low-pass</td>
<td></td>
<td>0.09</td>
<td>40</td>
</tr>
<tr>
<td>Fixed training</td>
<td>Low-pass</td>
<td>0.96</td>
<td>7</td>
</tr>
<tr>
<td>TDNN architecture C</td>
<td>High-pass</td>
<td>0.85</td>
<td>15</td>
</tr>
<tr>
<td>Low-pass</td>
<td></td>
<td>0.18</td>
<td>37</td>
</tr>
</tbody>
</table>

D. Localization performance of the TDNN with architecture B

The localization performance of the TDNN with architecture B is shown in Fig. 10. The localization is good in both the broadband and high-pass conditions and poorer in the low-pass condition. Compared to the subject’s localization, there was less aggregation of responses to a particular region in the low-pass condition. A quantitative comparison

FIG. 9. The localization responses of the TDNN with architecture A is shown in quadrature view. Other details as in Fig. 8. (a) Broadband localiza-

FIG. 10. The localization responses of the TDNN with architecture B is shown in quadrature view. The first hidden layer of this network is com-

(b) High-pass localization.

(c) Low-pass localization.

(b) High-pass localization.

(c) Low-pass localization.
of the network’s performance with that of the human subject is also given in Table VII. Most importantly, compared to the TDNN with architecture A, this TDNN was able to robustly incorporate temporal information into the localization process across the different bandpass conditions. Thus the frequency division imposed upon this network’s computational architecture improved the network’s localization performance.

E. Localization performance of the TDNN with architecture C

The TDNN with architecture C had a more realistic tonotopic structure (see Sec. III D) and was trained separately on the two different sound arrangements described in Sec. IV B. The results obtained when this TDNN was trained equally on the broadband, high-pass, and low-pass sounds are shown in Fig. 11. The localization performance was good for both the broadband and high-pass sound conditions and poorer for the low-pass condition, which produced many front-back errors. More interesting, perhaps, is the network’s localization performance when trained on sounds with random center frequencies and bandwidths. Not only did such training better approximate listening conditions with variable sounds, but it also provided a more substantial investigation into the network’s localization abilities. As discussed above, when randomly bandpassed sounds were presented during training, the learning task becomes considerably more complicated. The network was exposed to more than just three bandpass sound conditions during training. Nonetheless, this model produced localization performance across conditions that compared reasonably well with the localization of the human subject (see Fig. 12). Compared to the responses of the other TDNN models, this TDNN’s low-pass localization performance shows a closer resemblance to that of the subject; many of the localization responses were pulled downward and toward the front. A quantitative description of the performance of both of the networks described in this section are given in Table VII.

The localization performance of both the MLP and the TDNN with architecture C have demonstrated similarities with the human subject’s performance across the three different bandpass conditions (as discussed in previous sections). Nonetheless, the TDNN possesses the added advantage of incorporating temporal information along with the spectral information during the localization process. This additional and substantial realism constraint supplied the motivating factor for using the TDNN with architecture C in the continuing experiments that follow. These experiments probed the localization process using sounds with only restricted high frequencies and varying level.

F. Restricted high-frequency sound localization

Because the training data for TDNN (architecture C) were not fixed to any particular frequency region, it was reasonable to expect the network’s performance to generalize, for example, to a restricted high-frequency condition. Therefore the subject’s sound localization ability was further tested in the free-field using sounds bandpassed between 7637 and 13264 Hz. The subject’s localization results for a single test with 76 trial locations are shown in Fig. 13(a) and can be contrasted with that for the TDNN, shown in Fig. 13(b). The localization performance of the TDNN again shows qualitative similarities to that of the subject, with both demonstrating a clear aggregation of responses to restricted spatial regions in the lower hemisphere. A quantitative comparison of the TDNN’s localization performance with that of the human subject is given in Table VII. The localization
performance of both the subject and the TDNN, as shown here, are contrasted below with that of a matched filtering model.

G. Sound localization at varying sound levels

The sound level of the sounds used to train the networks were set and controlled via a parameter in the Auditory Image Model. The training of the TDNNs was performed in such a fashion, that for any particular location in space, the sound level did not vary by more than 1 dB SPL during repeated presentations of the sound. It was of interest to examine the robustness of the model’s localization performance under the influence of varying sound level. Thus, the localization performance of the TDNN (architecture C) trained on sounds with random center frequencies and bandwidth was examined, using a broadband sound source, across a range of sound levels varying from 60 to 80 dB SPL. The sound level used for training was approximately 67 dB SPL. It was observed that the transfer function of the inner hair cell model was essentially linear for this range of sound levels (within AIM, a medium spontaneous-rate auditory nerve fiber was used with a threshold shift of 0 dB, see Giguère and Woodland, 1994). The performance of the network was measured in terms of the spherical correlation coefficient between the target and response locations and is shown in Fig. 14. The spherical correlation coefficient peaked, of course, at the training sound level of 67 dB SPL and remained above 0.8 between 60 and 75 dB SPL. Angular errors corresponding with the correlation of 0.8 were typically between 10 and 15 degrees of spherical angle. In other words, a graceful degradation in localization performance was demonstrated over a range in sound level of 15 dB.

The same TDNN, as above, was also tested on broadband sounds, 10 dB louder in one ear than the other. The results of these tests are shown in Fig. 15 and clearly illustrate that the localization responses were pulled toward the side with the louder sound [Figs. 15(a), (b)]. The localization performance of the model was much improved when the sound level to both ears was increased [Fig. 15(c)]. While the magnitude of the effects shown here are certainly not human-like (see Wightman and Kistler, 1997), such behavior suggests that interaural level difference cues were a prominent and constant feature of the data that conferred some measure of robustness (in the model) to sound level variations.
H. TDNN encoding of temporal information

It has been argued above that the TDNNs encode temporal information as well as spectral information. Direct evidence for this fact is indicated in Fig. 16 where the weights for the first hidden layer of the TDNN with architecture C (trained on sounds with random center frequencies and bandwidth) are displayed. The weights for the low-frequency neurons clearly demonstrate a spectro-temporal structure while the weights for the high-frequency neurons indicate a constant spectral structure across time. This is in accordance with what one would expect given both the information provided via the auditory nerve response and the basic acoustics of sound diffraction around the head (Lord Rayleigh, 1907).

VII. DISCUSSION

A. Tonotopic processing from a computational viewpoint

As Secs. VIC–VIE clearly demonstrate, a network structure with frequency division, one that divides the input frequency channels into different groups and then progressively interconnects the neurons in the higher layers across frequency, was more robust in its localization performance across sounds with variable center frequency and bandwidth than a simple fully connected network. There are a few likely explanations for this observation. One line of reasoning argues that it may be easier for the tonotopic network to prevent a narrow band of frequency channels from dominating the localization computation across the entire set of sound stimuli. It was observed, in data not shown, that the low-frequency neurons required more training cycles to fully develop their weight structure than the high-frequency neurons. This observation is in accordance with the fact that these neurons demonstrate both a spectral and temporal structure in their connection weights (as shown in Fig. 16). The mismatch in training rates provides at least one source of conflict between the low- and high-frequency neurons. In the case of the TDNN with architecture A, in which each neuron within the first layer received input across the entire frequency range, the information within the different frequency channels may have destructively interfered with one another. The result was poor localization performance (Fig. 9).

Similar to the line of reasoning above, it can be argued that the frequency division imposed upon the network structure more evenly spreads the computation among the neurons in the different frequency groups and thus produces a more uniform computational structure. Evidence in line with this argument is the fact that for any single bandpass condition, it was found that the TDNN did not require frequency division within its architecture to produce quality solutions when trained only on these bandpassed sounds. In other words, it is the combination of frequency division and varying sound stimuli that encouraged the network to develop meaningful connections for all frequencies. The notion of a more uniform computational structure, as just described, is similar to two other ideas in the neural network literature. One is the notion of fault-tolerance in neural networks, an idea well explained by Neti et al. (1992) in their neural network models of the sound localization of the cat. Neti et al.
do not argue specifically for the connection between a tonotopic structure and fault-tolerance. However, they did develop a novel learning algorithm that minimizes a network’s error subject to the constraint that the increase in the error caused by removing a neuron from a computational layer of the network is uniformly small for all neurons within that hidden layer. To the extent that there is some overlap between spreading the computation across frequency and spreading the computation across neurons, one may encourage the other. It has also been argued (Shinn-Cunningham, 1998) that competitive learning with structural inhibition may allow the TDNNs to train appropriately across different bandpass conditions without requiring a tonotopic architecture. Such an argument assumes that competition would assist in distributing the computation across different neurons and supposedly, therefore, across frequency. Nonetheless, the network structure with frequency division, as demonstrated here, provided sufficient means for robustly training the TDNNs across the different bandpass conditions and thus played an important role in the computational process.

B. Matched filtering and sound localization

A number of previous sound localization models have used a relatively straightforward matched filter or template matching analysis (e.g., Middlebrooks, 1992; Nandy et al., 1993; Macpherson, 1997). In such cases, the ITD and spectrum of a given input sound is commonly cross-correlated with the ITD and spectrum of an entire database of sounds for which the location is known. The location with the highest correlation is then chosen as the optimal source location. In other words, matched filtering analysis can be seen as an optimal detection algorithm that, in the case of sound localization, assumes the system possesses a database of sounds, similar to the incoming sound, with which it has previously stored perfect knowledge of the source location with perfect resolution. The system then simply chooses the “best match.”

Matched filtering analysis was performed on the sounds bandpassed 7637–13264 Hz and the results are shown in Fig. 17. The matched filtering clearly localizes the sounds much better than the subject or the TDNN model as shown previously in Figs. 13(a)–(b) (a quantitative description of the performance is given in Table VIII). The matched filtering model used the same number of cochlear channels as the TDNNs and therefore contained the same inherent spectral resolution. This spectral resolution (31 cochlear channels) is certainly less than the spectral resolution of the human cochlea. This shows that although there was sufficient information to localize the sounds from the point of view of matched filtering, neither the human nor TDNN demonstrated such ability in their localization performance. This has important implications for understanding the TDNN localization responses shown in Fig. 13(b). Mathematically, a noncompetitive neural network, such as the TDNN, can learn to perform matched filtering. This can be seen from the formulas for the cross-correlation analysis and the output of a neuron in the neural network, respectively,

\[
corr_j = \sum_{i=1}^{n} \text{input}_{i}, \text{template}_{i},
\]

\[
\text{neuron}_j = f\left(\sum_{i=1}^{n} \text{input}_i, \text{weight}_i + \text{bias}_j\right),
\]

where corr is the cross-correlation result corresponding to the location labeled by , input is the /th spectral component of the input sound, template is the /th spectral component of the sound template corresponding to location , neuron is...
the output value of the neuron coding location $j$, weight $i_j$ is the network weight connecting the $i$th spectral component of the input sound with neuron $j$, bias $j$ is the constant network bias term for neuron $j$, and $f$ is the network squashing function. In order for the TDNN to localize similarly to the matched filtering model, the network weights corresponding to a given location need to assume the form of the filter template for that location. As all of the training sounds were flat-spectrum, the TDNN received no ambiguity as far as the source spectrum was concerned. Thus it is likely that the difference in the distribution of localization responses in Fig. 13(b), as compared with that in Fig. 17, has been encouraged by using training sounds with random center frequency and bandwidth. The results suggest that training with widely varying sound stimuli makes it difficult for the TDNN to achieve results comparable with matched filtering and that this may also apply to human sound localization performance. If so, this would provide a partial explanation as to why the human localization performance is not optimal from a matched filtering standpoint.

VIII. CONCLUSIONS

This study portrays a neural system identification model that was readily used to explore the computational process being performed by the human auditory system as it localizes sounds in space. During the development of the model, three physiological constraints were imposed upon the modeling process: (1) a TDNN model was used to incorporate the important role of spectral and temporal processing in the auditory nervous system, (2) a tonotopic structure was added to the network, (3) the training sounds contained randomly varying center frequencies and bandwidths. Not only did these constraints allow for a biologically plausible model, but more importantly they provided for an increased understanding of the role that these biological constraints play in the localization process. It was found that frequency division provided an essential ingredient for obtaining meaningful results across the different bandpass sound conditions. In particular, it allowed for the meaningful incorporation of temporal information into the low-frequency neurons of the network. In addition, it was argued that the tonotopic network structure more evenly distributed the computation over the hidden layer neurons, each of whose input was restricted in frequency. In this fashion, no particular frequency band could dominate the localization computation. As for the network training with variable sounds, it was assumed that this was a more biologically plausible listening condition than one in which the sounds were always broadband, though it makes the learning task more difficult for the network. The degradation in localization performance caused by training

<table>
<thead>
<tr>
<th>Group 10</th>
<th>LEFT</th>
<th>RIGHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIG. 16. The weights of the neurons in the first hidden layer are shown for the TDNN with architecture C (trained using sounds with random center-frequency and bandwidth). The weights of a given neuron are plotted in a block using gray-scale coloring. Time runs horizontally along the block and different frequency channels run vertically.

FIG. 17. The localization performance of the matched filtering algorithm is shown for the restricted high-frequency condition consisting of sounds bandpassed 7637–13264 Hz. The localization accuracy surpasses that of subject X and the TDNN shown in Fig. 13.

<table>
<thead>
<tr>
<th>Spherical correlation coefficient</th>
<th>Mean spherical angle error</th>
<th>Percentage front-back error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject X</td>
<td>-0.03</td>
<td>63</td>
</tr>
<tr>
<td>TDNN</td>
<td>0.03</td>
<td>68</td>
</tr>
<tr>
<td>Matched filtering</td>
<td>0.84</td>
<td>13</td>
</tr>
</tbody>
</table>

TABLE VIII. Comparison of the localization performance results for the subject, the TDNN with architecture C (trained using sounds with random center frequency and bandwidth), and the matched filtering model using a sound stimulus bandpassed 7637–13264 Hz.
with variable sounds was compared with human performance. The similarities obtained give some indication that this variability may partially explain why human sound localization is not optimal in terms of a simple matched filtering approach. In addition, the graceful degradation in localization performance with sound level variations was demonstrated. Finally, the modeling results obtained with the tonotopic network give some indication that the localization of arbitrarily bandpassed sounds can be related to external ear acoustics. However, given the discrepancies between the model and human localization performance, it may well be profitable to examine each cue to location individually in terms of its role or weighting in the localization judgment. This information may then inform future localization models with regard to the synthesis procedure that the auditory system must be using to combine information across the different cues to produce a coherent judgment of sound location.

ACKNOWLEDGMENTS

The authors would like to thank André van Schaik for his most valuable comments and insights and Marwan Jabri for his support. This research was supported by the Dora Lush Biomedical Postgraduate Scholarship to C.J. and the Australian Research Council (ARC Grant Nos. A49530190, A49700049, A79905421). The authors would also like to thank Pierre Zakarauskas and two anonymous referees for their valuable comments on an earlier version of this manuscript.

1 Any encoding of directions on a sphere that employs only two spatial variables must always be discontinuous because the surface of a sphere is not homeomorphic to the plane (see Guillemin and Pollack, 1974).

2 The time-delay networks progressively integrate more temporal information in the higher layers (in this case, up to about 8 ms).

3 It has been suggested byZipper (1992) that empirical evidence indicates that certain similarity classes exist in neural network solutions, although this has not been mathematically proven. Thus, it may be that for problems that are sufficiently “smooth” and networks that do not have too many hidden units, much the same computation is being performed.

4Target locations with an azimuthal angle within five degrees of the interaural axis were excluded from the front-back error analysis.

5These particular frequencies relate to the center frequencies of the model’s cochlear filters.

6The matched filtering analysis used templates derived from flat-spectrum noise bandpassed between 7637 and 13264 Hz, while some previous template-matching models have used templates derived from flat-spectrum, broadband noises. In other words, in this work the spectral analysis was constrained to the frequency region containing signal energy. There is no evidence suggesting that the auditory system performs spectral shape analysis in frequency regions with no signal energy. It is rather more likely that the large spectral contrast at the low-frequency band-edge may influence human sound localization performance, but will not bias the matched filtering model because of the within-band spectral assumption.


