

1992-02

A Neural Network for Synthesizing the Pitch of an Acoustic Source

<https://hdl.handle.net/2144/2089>

"Downloaded from OpenBU. Boston University's institutional repository."

**A NEURAL NETWORK FOR SYNTHESIZING
THE PITCH OF AN ACOUSTIC SOURCE**

Michael A. Cohen, Stephen Grossberg
and Lonce Wyse

February, 1992

Technical Report CAS/CNS-92-009

Permission to copy without fee all or part of this material is granted provided that: 1. the copies are not made or distributed for direct commercial advantage, 2. the report title, author, document number, and release date appear, and notice is given that copying is by permission of the BOSTON UNIVERSITY CENTER FOR ADAPTIVE SYSTEMS AND DEPARTMENT OF COGNITIVE AND NEURAL SYSTEMS. To copy otherwise, or to republish, requires a fee and/or special permission.

Copyright © 1992

Boston University Center for Adaptive Systems and
Department of Cognitive and Neural Systems
111 Cummington Street
Boston, MA 02215

A Neural Network for Synthesizing the Pitch of an Acoustic Source

Michael A. Cohen*, Stephen Grossberg† and Lonce Wyse‡

Center for Adaptive Systems and the Department of Cognitive and Neural Systems
Boston University, 111 Cummington St., Boston MA 02215

Abstract

This article describes a neural network model capable of generating a spatial representation of the pitch of an acoustic source. Pitch is one of several auditory percepts used by humans to separate multiple sound sources in the environment from each other. The model provides a neural instantiation of a type of “harmonic sieve”. It is capable of quantitatively simulating a large body of psychoacoustical data, including new data on octave shift perception.

1 Background and Model

A fundamental problem of auditory and speech perception is the identification and separation of multiple acoustic sources. Such a process enables human listeners to detect and recognize the contents of discriminable auditory streams, in a process called auditory scene analysis by Bregman (1990). The process utilizes a variety of cues including synchrony, pitch, and localization information to assign acoustic components to the appropriate auditory stream. The present article describes a neural network model for generating a spatial representation for the pitch of an acoustic source that can be naturally imbedded in an architecture for signal separation.

The current model is a type of “pattern matching” model, a class that also includes the pitch models of Goldstein (1973) and Wightman (1973). The input to the pitch detecting module is a spectral representation discussed below. Each possible pitch samples regions of the spectrum with a sampling period equal to the pitch frequency. That is, a region around nf_0 , for integers n and fundamental frequency f_0 , contributes to the strength of the pitch percept at frequency f_0 . The weighting function for the region is Gaussian and symmetric in log frequency space (Figure 1), causing the resolution of the filter to scale with frequency.

The model matches significant pitch perception data for reasons similar to those of the pattern matching models of Goldstein and of Wightman, but has significantly different implications for its neural instantiation. The latter two models have a close mathematical relationship which was demonstrated by de Boer (1976). In each of these theories, evidence is accumulated for a particular pitch percept by matching a template to a spectral representation of the signal. A key component in each is a filter whose bandwidth scales with its center frequency, which spreads or randomizes the effect of a component across frequency in the spectral representation.

*Supported in part by the AFOSR (90-0128).

†Supported in part by the AFOSR (90-0175), DARPA (AFOSR 90-0083) and the NSF (IRI 90-24877).

‡Supported in part by the American Society for Engineering Education.

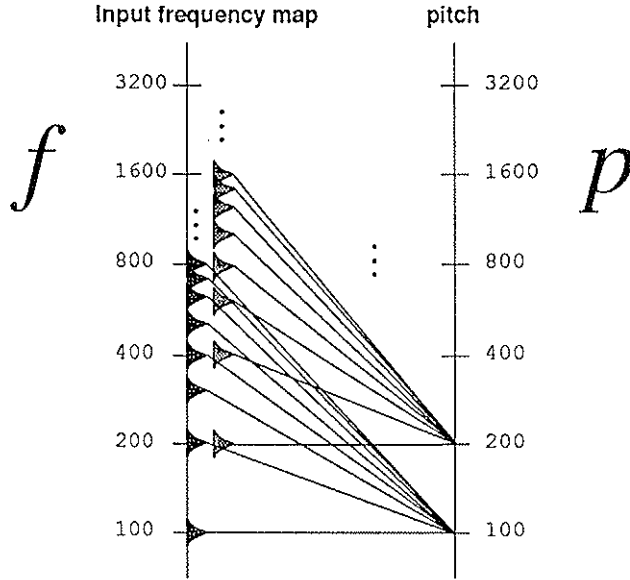


Figure 1: *Each pitch p samples the input frequency map around integer multiples of the corresponding pitch. The kernels are Gaussian and symmetric in \log_2 coordinates.*

A key difference between the Wightman model and the Goldstein model is that Wightman's is a deterministic model which produces a strength of activation for every pitch. Wightman's bandwidth-scaling filters are the peripheral auditory filters, modeled with a triangular shape in log-log coordinates, and are intended to approximate the resolving powers of rate place coding performed by basilar membrane mechanics. In Goldstein's model, the frequency scaling function is a probability density function. Each component is passed to the central spectrum through a normally distributed random error generator. Wightman performs a cosine Fourier transform on the smeared spectral representation yielding the pitch activation function. Goldstein finds the best match in least squares of the noisy spectrum and a perfectly harmonic template yielding the most likely pitch. This model can also be extended to produce a probability density function across all pitches.

The model described herein also uses a frequency scaling kernel that spreads the effect of each component. This spreading is built into the sieve used to generate pitch strength. In addition to the Gaussian sieve weighting function, the model incorporates data which shows that higher harmonics have a lesser effect on pitch than lower harmonics (Ritsma 1962, 1963, Plomp 1967). We have used a simple function of harmonic number that decreases linearly with harmonic number. No absolute frequency region is given preferential weighting. This weighting function is important not just to capture data on the ability of different regions to affect small pitch shifts as they are mistuned, but it is critical for controlling perceived octave shifts in response to ambiguous stimuli as shown below. Thus the steady-state equation that instantiates the network in Figure 1 for strength S_p of each pitch p is:

$$S_p = \sum_n \sum_f h(n) I(f) \exp\left(-\frac{1}{2} \frac{(\log_2(np) - \log_2(f))^2}{\sigma^2}\right)$$

$$h(n) = \begin{cases} 1 - \lambda n & \text{for } \lambda n < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $I(f)$ is the strength of the spectrum at frequency f , np is the frequency of the n th harmonic of the pitch p , and $h(n)$ is the harmonic number weighting function, linear with slope $-\lambda$. The “best fitting fundamental” is taken to be the one with the strongest activation, but the strength function bears a striking resemblance to the probability density function derived by the Goldstein model. A winner-take-all operation, implemented by a shunting on-center off-surround network (Grossberg 1973, 1988), is used to select the maximally activated pitch.

The model is fed a spectral pattern as input, but the method of computing the input pattern is not constrained by the pitch detector. Any of the many techniques explored by other modelers (Scheffers, 1983; Terhardt, 1974; Young and Sachs, 1979) which produces reasonably resolved components could be used. While the model takes a place code as input, a time-place model of the auditory periphery is implied by the need for a spectral representation that is approximately invariant over a wide dynamic range and under different noise conditions (Young and Sachs, 1979; Srulovicz and Goldstein, 1978).

2 Comparisons with data

When harmonic components ($f_n = nf_0, n = 1, \dots$) are all shifted by a constant difference so that they maintain their spacing of f_0 , the pitch shifts at a slower rate than the harmonics (Schouten 1943, 1962; Patterson and Wightman 1976). The typical data reported show an ambiguous region at shift values of $f_0/2$ where the perceived pitch suddenly jumps down to below the level of f_0 and begins increasing again toward f_0 as the lowest component again approaches an integer multiple of f_0 (Figure 2). The model shows a close correspondence with the data. The pitch shifts slower than the harmonics because the width of the Gaussian kernels scales with frequency. Because of the frequency scaling, as the harmonics shift, the higher harmonics move through the Gaussian kernels of a particular pitch much slower than the lower harmonics. Thus, the pitch shifts, and the maximum pitch strength weakens as the different components become centered on harmonics of nearby, but different, pitches.

Much of the pitch shift data has been gathered by focusing the experimental subject’s attention on a narrow pitch region centered at f_0 , and has thus neglected the true extent of the ambiguity of the pitch sensation in the ambiguous region. Modelers have similarly restricted their models to pitch decisions in a small region around the fundamental f_0 . In fact, as Schouten (1943) showed, the distribution of pitch matches is multi-modal with the various modes being clearly separated. Several of the modes are near f_0 , but several are further away and have not shown up in the data due to the experimental methodology. The ambiguous region of the harmonic shift is characterized by the components being near the frequencies $f_n = f_0(1/2 + n)$ which can be written as $mf_0/2$ for odd integer m . That is, the ambiguous region is where the components are all near odd harmonics of $f_0/2$. Raatgever and Bilsen (1991) have recently demonstrated a clear tendency of subjects to report hearing $f_0/2$ in the ambiguous region. Furthermore, the tendency shows a systematic decline as fewer and fewer low harmonics are contained in the stimulus. The model predicts both the octave shift in the ambiguous region as well as the systematic way it changes with the absence of lower harmonics (Figure 3).

As a single component in a harmonic complex is mistuned, the perceived pitch begins to shift

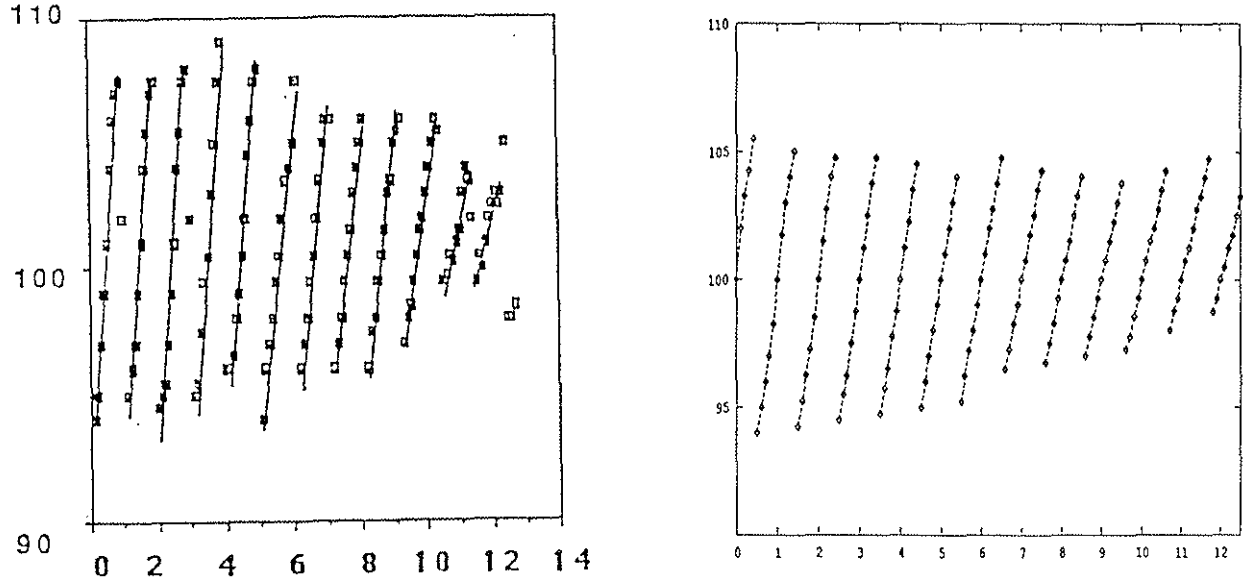


Figure 2: *Pitch shift as a function of the lowest component's harmonic number. On the left, data from Patterson and Whightman, 1976 (reprinted with permission). On the right, maximally activated pitch produced by the model.*

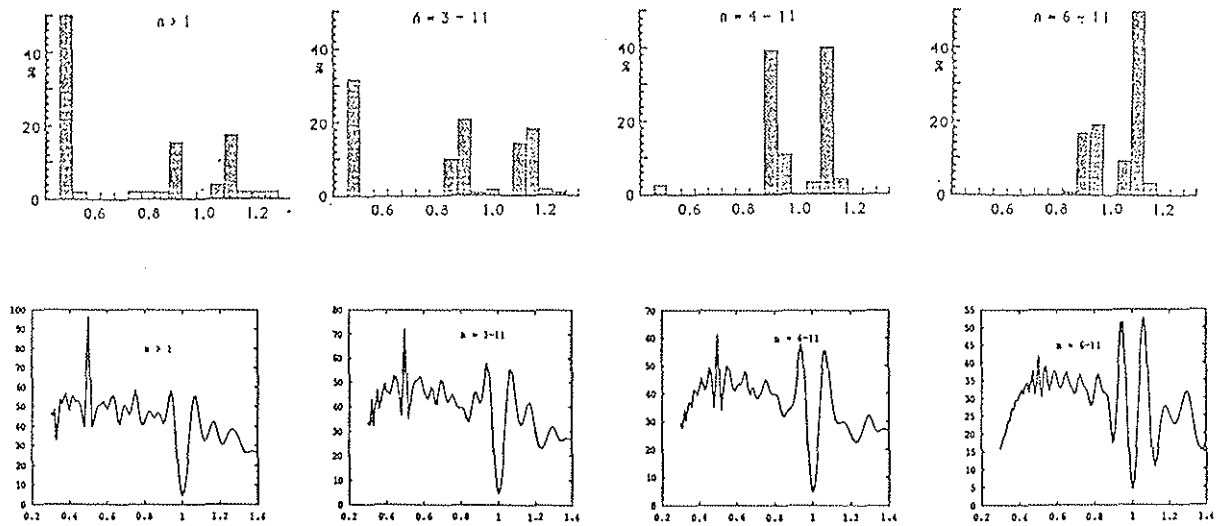


Figure 3: Top Row: Distribution of pitch matches as a function of f_{match}/f_0 (Raatgever and Bilsen, 1991. Reprinted with permission). As the lower harmonics ($n = 1..5$) are removed, the tendency to hear the pitch an octave below f_0 disappears. Bottom Row: Model pitch strength (in arbitrary units) as a function of f_{pitch}/f_0 .

at first in the same direction as the component. As the component is mistuned beyond 3% of its original frequency, its effect on the pitch begins to diminish and the pitch shifts back toward its original f_0 . When the component is mistuned by roughly 8% of its original frequency, its effect on the pitch is no longer perceptible. Moore et al suggested that if a harmonic sieve is operating, one possible explanation of these data is that a component does not fall through the sieve in an all-or-none fashion. In our model, the reason for the effect is the Gaussian shape of the weighting function.

The model also predicts the repetition pitch phenomena which are created by delaying wide band noise and adding it to itself (Bilsen, 1966). If the noise is added to itself with a positive sign, the pitch is perceived as the reciprocal of the delay. In our model, this happens because the amplitude spectrum for such a signal has peaks that are separated by this amount. If the noise is added to itself with a negative sign, then the pitch is ambiguous. In our model, this is due to peaks in the amplitude spectrum which are still separated by the reciprocal of the delay, but are shifted by half this amount. That is, the spectral peaks are at the same locations that characterized the ambiguous region discussed above.

3 Conclusion

The present model bases its pitch representation on a spectral representation of the components. Ongoing research will extend the model to capture phase sensitive effects in pitch perception. Similarly, AM modulated noise, which has a flat amplitude spectrum independent of modulation frequency, but can nevertheless produce a weak pitch percept, is not explained by the model in its current form. We point out, however, that pattern matching models of pitch perception do not mandate that the spectrum upon which they operate is the spectrum of the sound source, but rather an internal representation generated along a pathway that contains significant nonlinearities in both the mechanical and neural processing. Thus, AM modulated noise might produce an internal spectral representation with enough shape for a pattern matching module to produce a weak pitch percept.

The model produces as output a strength value across a spatial representation of pitch, rather than merely producing the frequency of the most likely pitch, such as models that base the decision on the fine structure of a temporal waveform. Such a representation is important not only because it can provide an explanation for data on responses to ambiguous stimuli, but also because the representation forms part of the dynamics of the system in which it is embedded. For example, if attentional factors are used to prime a particular frequency region, then the spatial representation plays an important role in determining the ensuing pitch percept. Such a spatial representation also has a more direct neural interpretation than do models that make use of Fourier transforms or autocorrelations. Most importantly, the model can be embedded in a larger architecture which uses pitch as one of many cues to group the components of different sound sources and to separate these sources from one another in the auditory scene.

References

- [1] Bilsen, F.A. (1966) Repetition Pitch: monaural interaction of a sound with the repetition of the same, but phase shifted, sound. *Acustica* 17, 265-300.

- [2] Boer, E. de (1976). Pitch Theories unified. In E.F. Evans and J.P. Wilson (Eds.), **Psychophysics and Physiology of Hearing**. Academic, London.
- [3] Bregman, A. (1990) **Auditory Scene Analysis**. MIT Press. Cambridge, Massachusetts.
- [4] Goldstein, Julius L. (1973) An optimum processor theory for the central formation of the pitch of complex tones. *J. Acoust. Soc. Am.* **54**(6), 1496-1516.
- [5] Grossberg, S. (1973). Contour enhancement, short-term-memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, **52**, 217-257.
- [6] Grossberg, S. (1988). Nonlinear Neural Networks: Principles, Mechanisms, and Architectures. *Neural Networks* **1**(1).
- [7] Meddis, R. and M.J. Hewitt (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *J. Acoust. Soc. Am.* **89**(6), 2866-2882.
- [8] Moore, B.C.J. and B.R. Glasberg. (1985) Relative dominance of individual partials in determining the pitch of complex tones. *J. Acoust. Soc. Am.* **77**(5)
- [9] Patterson, R., and F. L. Wightman (1976). Residue pitch as a function of component spacing. *J. Acoust. Soc. Am.* **59**(6) 1450-1459.
- [10] Plomp, R. (1967) Pitch of Complex Tones *J. Acoust. Soc. Am.* **41**(6) 1526-1533.
- [11] Raatgever, J. and F.A. Bilsen (1991). The pitch of anharmonic comb filtered noise reconsidered. Proceedings of the 9th International Symposium on Hearing, Carcans, France.
- [12] Ritsma, R. J. (1962). Existence region of the tonal residue. I. *J. Acoust. Soc. Am.* **34**, 1224-1229.
- [13] Ritsma, R. J. (1963). Existence region of the tonal residue. II. *J. Acoust. Soc. Am.* **35**, 1241-1245.
- [14] Scheffers, M.T.M. (1983). Simulation of auditory analysis of pitch: An elaboration on the DWS pitch meter. *J. Acoust. Soc. Am.* **74**(6) 1716-1725.
- [15] Schouten, J.F., J.R. Ritsma, and B.L. Cardozo (1962). Pitch of the Residue. *J. Acoust. Soc. Am.* **34**(8), 1418-1424.
- [16] Srulovicz P. and J.J. Goldstein (1983). A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum. *J. Acoust. Soc. Am.* **73**(4) 1266-1276.
- [17] Terhardt, E. (1974) Pitch, consonance and harmony. *J. Acoust. Soc. Am.* **55**(5) 1061-1069.
- [18] Wightman, F.L. (1973). The pattern-transformation model of pitch. *J. Acoust. Soc. Am.* **54**(2) 407-416.