数理解析研究所講究録
第828巻 1993年 1-10

聴神経の時系列発火に基づく音調性認識のニューラルネットモデル

京都大学工学部　喜多　一

1. はじめに

人間の聴覚系では外耳で捉えられた音が内耳にある蝸牛の機械的な構造により周波数分析され，蝸牛内の有毛細胞が機械振動を電気信号に変換して聴神経を刺激し，中枢へと伝えられる．個々の聴神経は音の特定の周波数成分に極めて選択的に応答する．すなわち，音の情報は周波数に選択的に応答する聴神経の位置により表現されている．このような情報表現を手がかりに聴覚における種々の感覚を説明する理論は 'place theory' と呼ばれている．一方，聴神経の発火は約 5kHz までは音の特定の位相に同期して生じることが知られている．このことは聴覚系における音声情報の処理が聴神経発火の時系列パターンに基づいて行われている可能性をも示唆する．聴神経発火の時系列パターンを手がかりに聴覚現象を説明しようとする理論は 'temporal theory' と呼ばれている [1].

音の周波数に関連して，人間には 2 種類の感覚がある．一つは周波数の上昇に対して単調に変化する感覚であり，もう一つは 1 オクターブ離れた 2 音に対して類似性を感じる感覚である．後者は音楽において極めて重要な役割を演じるものである．大串は前者を「かん高さ (tone height)」，後者を「音調性 (tonality)」と呼んだ [3, 4]．音調性に関しては，以下のような生理学的あるいは心理学的知見が同感覚と先に述べた聴覚における temporal theory との関連性を示唆するものとして注目される：

1. かん高さに比べ，音調性を感じる周波数の上限はより低く，約 5kHz 以上の帯域においてはこの感覚は消失する．そして，この音調性を感じる限界の周波数と聴神経の位相同期発火の限界周波数はよく似た値をとる．
2. 音響心理実験で計測される「主観的なオクターブ」，すなわち 2:1 前後の周波数比において心理的に最も類似性を強く感じる周波数比は「物理的なオクターブ（周波数比が厳密に 2:1）」に比べやや大きい．

これらの知見に基づいて大串は音調性が聴神経発火の時系列パターンに由来するという説を提案した [3]．しかしながら，これまでのところ，音聴性という感覚の神経メカニズムを示唆する生理学的知見は見当たらない．

本論文では，音聴性について聴覚の temporal theory に基づき，可能なメカニズムのひとつとして，ニューラルネットモデルを提案する．このモデルでは，音声信号は有毛細胞に相当する非線形フィルターにより変形されたのち，遅延線により空間パターンに変換される．そして，これが 2 次元の自己組織化特徴地図 (self-organizing feature map, 以下 SOFM と略称する) に入力される．計算機シミュレーションにより，本モデルではオクターブ周期性をもったトノトピック・マップ（tonotopic map, 特定の周波数に応答する神経細胞が周波数に対して連続的に空間配置された神経地図）の形成が行われることが示される．

2. ニューラルネットワークモデル

2.1 モデルの構造

本論文で提案するモデルは図 1 のようなもので，次のような構造を持つ：

- 入力された音声に，まず整流器特性を持つ非線形変換が施される．この特性は蝸牛における有毛細胞の特性 [1] をモデル化したものである．
- 次に時系列的信号が遅延線により空間的なパターンへと変換される．遅延線全体の遅延時間は対象とする信号の周波数帯域で数波程度に相当する時間を仮定する．また，遅延線上で信号は遅延にともなって減衰することも仮定する．
- 最後に空間パターンに展開された信号は 2 次元状に整列する 1 層のニューロン群に送られる．ただし，信号は遅延線の入り口部分に波形のピークが現われた時点でサンプリングされるものと仮定している．この層のニューロン群は Kohonen によって提案された自己組織化特徴地図の学習アルゴリズム [5] により教師なし学習を行う．

2.2 音声の内部表現

以下，簡単のため入力信号 x は周波数が f で単位振幅を持つ純音（正弦波）であると仮定する:

$$x(t,f) = \cos(2\pi f t) \tag{1}$$

ここで t は時間を表す．なお，波形を余弦関数で表すのは後に行うサンプリング後の表現を簡単にするためである．入力信号 x は蝸牛における有毛細胞に相当し，整流特性を持つ非線形変換 g を受ける：

$$y(t,f) = g(x(t,f)) \tag{2}$$

$$g(x) = \begin{cases} x^k & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases} \tag{3}$$

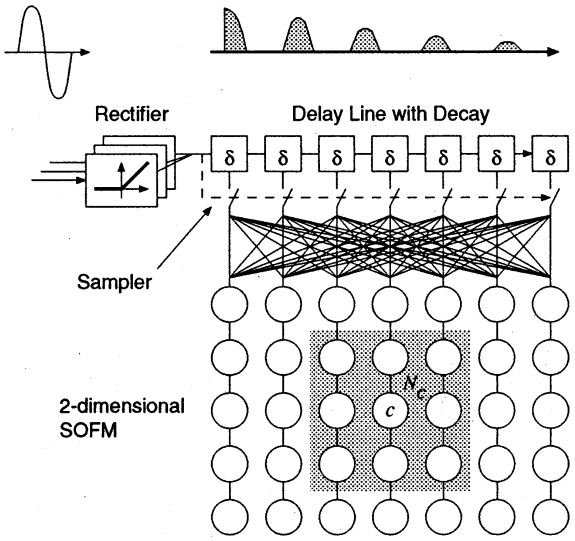

図 1: トノトピック・マップ形成のニューラルネットワークモデル

ここで k は正のパラメータである．k が大きければ，非線形変換 g は入力信号のピークを強調する．

次に，信号は 1 段あたりの遅延時間が δ で減衰率 r の遅延素子 $N-1$ 段からなる遅延線により空間的パターンにより変換される:

$$y_i(t,f) = r^i y(t-\delta i, f), \quad i = 0,\ldots,N-1 \tag{4}$$

ここで y_i は i 段目の遅延素子の出力である．

さらに，空間パターンに展開された信号は遅延線の入力端にそのピークが現われた時点でサンプリングされるものと仮定する:

$$y_i(f) = r^i g(\cos(2\pi f \delta i)), \quad i = 0,\ldots,N-1 \tag{5}$$

この空間パターン (5) が自己組織化特徴地図に提示される．

機能的な観点から上記の構造は次のような特徴を持つ:

- 整流器状の非線形特性 g は 1 オクターブ離れた純音間に類似性を与える．これは，1 オクターブ離れた純音間では低い方の音の正のピークが現われる時刻が高い方の音の正のピーク（の半数）が現われる時刻に一致しており，また変換前には 2 音で波形の正負が反転することのある負のピークが整流特性により除去されるため，2 音の変換後の波形の類似性が高まるためである．
- 減衰を伴う遅延線は周波数の近い純音間の類似性をもたらす．

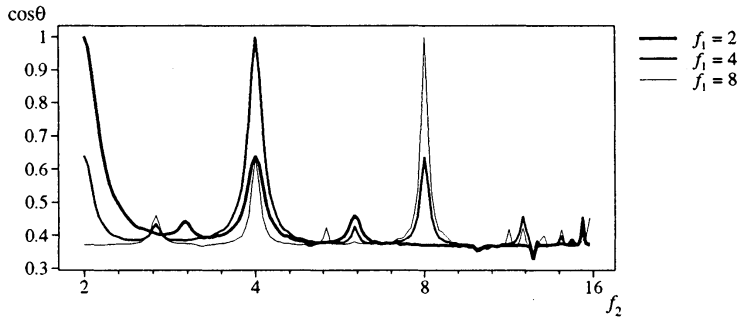

図 2: 遅延線上に表現された純音間の類似性.

- 波形のピークが遅延線の入力端に現われた時点で行なわれるサンプリングは音声の位相変化を吸収する.

このようにして変換された信号 (5) のトポロジカルな構造を検討してみる. 周波数 f_1 及び f_2 の 2 純音について上記のプロセスにより変換されたパターンをそれぞれ$\boldsymbol{y}_1 = (y_0(f_1), \ldots, y_N(f_1))$ 及び$\boldsymbol{y}_2 = (y_0(f_2), \ldots, y_N(f_2))$ とする. これらのパターン間の類似性を $\boldsymbol{y}_1$ と $\boldsymbol{y}_2$ の方向余弦

$$\cos\theta = \frac{\boldsymbol{y}_1 \cdot \boldsymbol{y}_2}{|\boldsymbol{y}_1||\boldsymbol{y}_2|} \tag{6}$$

で計る. 図 2 は f_1 を 2, 4 及び 8 に固定し, f_2 を 2 から 16 まで変化させたときの $\cos\theta$ の変化である. 用いたパラメータの値は $k = 2$, $N = 200$, $\delta = 1/50$ 及び $r = 0.99$ である. 同図から内部表現 (5) は, 周波数の近い純音間では類似性が高く, また 1 オクターブ離れた純音間でも類似性が高いという構造を持つことが分かる.

2.3 自己組織化特徴地図

自己組織化特徴地図 (SOFM) は Kohonen によって提案されたニューラルネットであり, 教師なし学習により, 入力信号の持つトポロジカルな構造を発火するユニット (ニューロン) の位置に反映するようなネットワークの結合重み (シナプス荷重) を形成するという特徴を持つ [5]. SOFM の動作及び学習規則は次に述べるものである.

まず, 変換された信号 $\boldsymbol{y}$ が SOFM に提示されると, 信号 $\boldsymbol{y}$ に最も近い結合重み $\boldsymbol{m}_c$ を持つユニット c が選択される. すなわち

$$c = \arg\min_j \| \boldsymbol{y} - \boldsymbol{m}_j \| \tag{7}$$

そしてネットワークの出力 z_j は次式で与えられる:

$$\begin{aligned} z_c &= 1 \\ z_j &= 0 \quad \text{for } j \neq c \end{aligned} \tag{8}$$

またSOFMの学習ルール，すなわち結合重みの変更ルールとしては次のものを用いる:

$$\begin{aligned} \boldsymbol{m}_j(T+1) &= \boldsymbol{m}_j(T) + \alpha(T)(\boldsymbol{y}(T) - \boldsymbol{m}_j(T)), \\ &\qquad \text{for } j \in N_c \\ \boldsymbol{m}_j(T+1) &= \boldsymbol{m}_j(T), \quad \text{otherwise} \end{aligned}$$

ここで T はパターンがSOFMに提示される毎に1づつ増加する時間のインデクスである．$\alpha(T)$ は学習係数と呼ばれ，時間とともに減少する正のパラメータである．また，N_c は選択されたユニット c の近傍に位置するユニットの集合である．

3. シミュレーション

3.1 実験 1

2節で述べたニューラルネットモデルを用いて，以下のような設定で計算機シミュレーションを行なった:

- 入力信号: 入力信号は振幅が1で周波数は2から16までの間を対数スケールで等分した144通りの正弦波である．
- 非線形変換と遅延線: 非線形変換 g と遅延線のパラメータは2.2節で用いたものと同じである．
- SOFM: 自己組織化特徴地図には 20×20 の正方形状に配列した400ユニットを用いた．近傍 N_c も正方形状のものを用いた．

学習回数を5000回，$\alpha(T)$ を学習回数に対して直線的に0まで減少する関数 $0.2(5000-T)/5000$ に設定して学習を行なった．シミュレーション結果を図3に示す．自己組織化学習により，同図を見れば，周波数に関する連続性とオクターブ周期性を合わせ持った螺旋上のトノトピック・マップがSOFMに形成されていることが分かる．このような構造は2.2節で見たように変換された信号 (5) が持つ位相構造を反映したものである．

3.2 実験 2

実験1では，入力信号を振幅1の純音に制限しており，また遅延線により空間パターンに展開された信号はピークが遅延線の入力端に現われるときに正確にサンプリングできるものとし

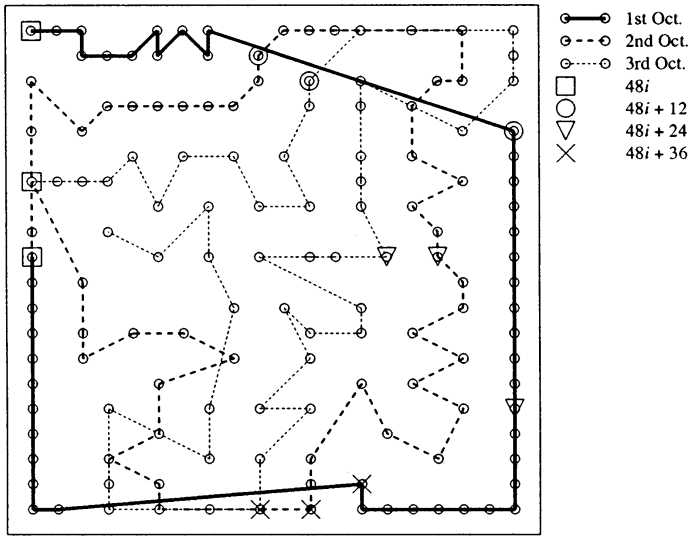

図 3: シミュレーション結果 1. 学習終了後のネットワークに周波数 2 から 16 までの純音を提示し，発火したユニットを示してある．□○▽×はそれぞれオクターブ内で相対的に同じ位置にある周波数の純音に応答したユニットを表す．

ていた．しかしながら，現実に我々が耳にする音では時事刻々その振幅，位相，周波数成分などが変化しており，先の実験の仮定は現実的ではない．実験 2 では学習に際してより複雑な信号を用い，SOFM におけるトノトピック・マップの形成について検討する．入力信号は次のように生成した:

- 入力信号は基本波のほかに 2 倍及び 3 倍の高調波成分からなる．
- 基本波の周波数は区間 $[1,8]$ から対数スケール上で一様乱数により選ぶ．
- 基本波の振幅は区間 $[2/3,1]$ 間で一様乱数により選ぶ．
- 2 倍音及び 3 倍音の基本波成分に対する相対振幅はそれぞれ区間 $[1,1/2]$ 及び $[1,1/3]$ 間で一様乱数により選ぶ．
- 各調波成分の位相は区間 $[0,2\pi]$ から一様乱数により選ぶ．

上記の手続きで生成した信号のうち時刻 0 での変位が 0.7 を超えるものを 1000 件選び，ネットワークの学習に用いた．学習回数は 2000 回，$\alpha(T)$ は $0.2(2000-T)/2000$ と設定して学習を行なった．

学習終了後，振幅 1 で周波数 1 ～ 8 の純音を入力してネットワークの応答を調べた．各純音に対して発火したユニットを図 4 に示す．同図を見れば実験 1 と同様に，オクターブ周期性を持ったトノトピック・マップが形成されていることが分かる．図 3 と比較すると，学習用のデータが複雑であることを反映して，螺旋状の構造はかなり歪んだものとなっている．

4. 考察

2 節で提案したモデルは音声情報が聴神経発火の時系列パターンによって表現され得る事に着目し，音調性というオクターブ類似性を持つ感覚が生じる起源を蝸牛にある有毛細胞の持つ整流器状の非線形特性に求めたものである．3 節では計算機シミュレーションにより提案したモデルがオクターブ周期性を持つトノトピック・マップを学習により自己組織的に獲得することを示した．本モデルの利点はオクターブ周期性を持つマップが純音のみを入力とする環境下でも獲得され得ることにある．これに対し，蝸牛における周波数分析とそれに伴い発火する聴神経の位置による情報表現に基づく理論 (place theory)[2] ではオクターブ周期性の獲得は入力となる音声の倍音構造に大きく依存してしまう．

次に提案したモデルの生理学的な妥当性と計算論的な問題点について若干の検討を加えておく．本モデルでは遅延線が重要な役割を果たすが，要求される遅延時間は例えば 1kHz 程度の音声に対して，数波分の時間，すなわち o(msec) 程度の時間である．これまで，聴覚系において，このような長さを持つ遅延線の存在は報告されていない．ただし，やはり聴神経発火の時系列パターンを用いて両耳への音の到達時間差を検出しているフクロウの音源定位メカニズム

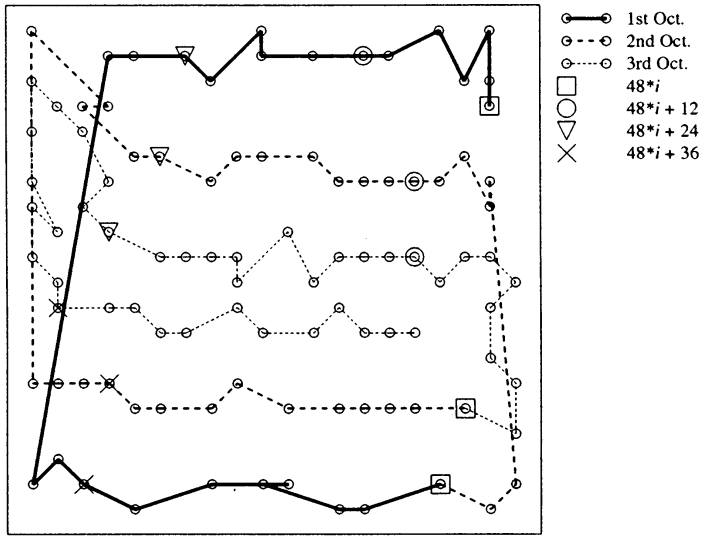

図 4: シミュレーション結果 2. 種々の振幅，位相，高調波成分を持つ信号で学習したネットワーク.

においては $o(100\mu\text{sec})$ 程度の遅延線の存在が報告されている [8].

また計算論的な問題としては，複合音，とくに非調波的な複合音に対して本モデルではトノトピック・マップの形成が困難になることが挙げられる．この問題点については，さらに検討を行なう必要があるが，関連すると考えられる音響心理学的な知見としては，2 音を同時に提示したときに知覚される心理的なオクターブ (harmonic octave) と 2 音を逐次的に提示したときに知覚される心理的なオクターブ (melodic octave) では特性が異なっているという報告 [6, 7] がある．これらのことを考慮すると，音調性は単に聴神経発火の時系列パターンや空間パターンのいずれか単独の情報表現に基づくのではなく，両表現を統合して知覚されいるのかもしれない．このような 2 種類の表現両方に基づく知覚メカニズムの理論を検討することも今後の課題である．

5. おわりに

本研究では「音調性」と呼ばれる音の周波数に関係し，オクターブ類似性を持つ聴覚について，その可能なメカニズムを考察した．本研究で提案したモデルは音声情報が聴神経発火の時系列パターンによって表現されうる事に着目し，音調性というオクターブ類似性を持つ感覚が生じる起源を有毛細胞の非線形特性に求めたものである．計算機シミュレーションにより，提案したモデルがオクターブ周期性を持つトノトピック・マップを自己組織的に形成することを示した．

最後に本研究にあたってご指導，ご討論を頂いた京都大学，西川 禕一教授ならびに京都市立芸術大学，大串健吾教授に感謝の意を表します．

参考文献

[1] J. O. Pickles: *An Introduction to the Physiology of Hearing, 2nd Ed.*, Academic Press (1988).

[2] E. Terhardt: "Pitch, consonance, and harmony," *J. Acoust. Soc. Am.* Vol. 55, pp. 1061-1069 (1974).

[3] K. Ohgushi: "The origin of tonality and a possible explanation of the octave enlargement phenomenon", *J. Acoust. Soc. Am.* Vol. 73, pp. 1694-1700 (1983).

[4] 大串:「複合音の高さの循環性とその応用」，電子通信学会論文誌，Vol. J67-A, pp. 423-430 (1984).

[5] T. Kohonen: *Self-Organization and Associative Memory, 2nd Ed.*, Springer-Verlag (1988).

[6] L. Demany and C. Semal: "Harmonic and melodic octave templates," *J. Acoust. Soc. Am.*

Vol. 88, pp. 2126-2135 (1990).

[7] L. Demany, C. Semal and R.P. Carlyon: "On the perceptual limits of octave harmony and their origin," *J. Acoust. Soc. Am.* Vol. 90, pp. 3019-3027 (1991).

[8] U.E. Sullivan and M. Konishi: "Neural map of interaural phase difference in the owl's brainstem," *Proc. Nat. Acd. Sci. USA*, Vol. 83, pp. 8400-8404 (1986).

ABSTRACT

A Neural Network Model of Tonality based on the Temporal Theory of Auditory Sensation

Hajime KITA

Dept. of Electrical Engineering, Kyoto University

Yoshida-Honmachi, Sakyo, Kyoto 606-01, JAPAN

The human being is equipped with two sorts of auditory sensation regarding the frequency of sound. The one is a sensation changing monotonically according to increase in the frequency of a tone. The other is a sensation having octave affinity. Ohgushi called the former one 'tone height' and the latter 'tonality'. In the cochlea, the sound is decomposed into its frequency components. At the same time, auditory nerves can follow the temporal change of the wave of sound up to about 5 kHz. Several psychoacoustic and physiological findings suggest existence of a close relationship between the sensation of tonality and such a temporal representation of sound in the auditory system.

In the present paper, the authors propose a neural network model which is to explain a possible mechanism of the sensation of tonality. The proposed model consists of three parts, i.e., a nonlinear transformation of the sound, conversion of the temporal signal into the spatial one by a delay line, and Kohonen's self-organizing feature map (SOFM). Simulation results show formation of tonotopic map having octave periodicity in the SOFM.