2E3-1

音声による子音カテゴリカル知覚の習得モデルに関する研究

Study on a model of learning for categorical perception of Voiced-Voiceless plosive consonants

宮澤 幸希* 白勢 彩子* 菊池 英明*
MIYAZAWA Kouki SHIROSE Ayako KIKUCHI Hideaki

* 早稲田大学人間科学学術院
Faculty of Human Sciences, Waseda University

Abstract: We examined a learning process of categorical perception of Voiced-Voiceless plosive consonants by using neural network model. In this study, *Self-Organizing Maps* (SOM) model learned values of Voice Onset-Time (VOT) of two consonants (/d/, /t/) and was verified a boundary value of two consonant categories. We referred to data of previous results of VOTs perception and production of American speakers and chinchilla as the input value of SOM. Our results demonstrated that SOM could categorize the values of VOT into two groups, /d/ and /t/, and the value of the categorical boundary showed a good fitness to human data appeared in previous studies.

1. 研究の目的

本研究は，人間の知覚メカニズムを手本とした音声言語理解技術を開発することを目的として，人間の知覚現象の一つである「子音カテゴリカル知覚」の習得過程をコンピュータモデルによって再現しようとするものである．

現在コンピュータによる音声言語理解は，大量の学習データに基づいた統計的なモデルを用いる方法が一般的であるが，状況や環境の変化に柔軟に対応できず不完全なものであると言わざるをえない．それに対して，人間は言語のある環境で生活すれば，誰でも高度な言語能力を獲得できる．

そこで，人間が行っている言語獲得のメカニズムを解明し，それを応用することによって，高度な適応能力を持った音声言語理解技術の開発を目指す．

なお，子音の同定に必要な音響的特徴を連続的に変化させて聞き取ると，連続的には知覚されず，変化が一定の範囲内にあれば同一の子音に知覚される．このような知覚様式をカテゴリカル知覚という[1].

人間の子音カテゴリカル知覚の習得メカニズムは，言語経験による学習と生得的な言語機構の両面からなると考えられている．この関係を解明するため，可能な限り生得的機構を除くことを仮定した学習モデルで子音カテゴリ境界の獲得を検証し，その結果を実際の人間の子音知覚境界と比較して検討した．

2. 自己組織化マップの利用

子音カテゴリカル知覚の習得をシミュレーションするために，行動主義心理学の強化学習，ニューラルネットワーク，統計的学習モデルを導入し，それぞれの機能を比較して予備的に考察した結果，コホネンの Self-Organizing Maps[3]（自己組織化マップ，以下 SOM）が子音カテゴリカル知覚の学習モデルに最も適していると判断した．

SOM は，大脳皮質における神経細胞の情報処理を参考に考案された自己組織化ニューラルネットワークの一種である．

実験によって，SOM には，

- 入力値の分布傾向を自動的に（教師信号や報酬なしに）分析し，カテゴリ分けする
- 一度も学習していない値でもどのカテゴリに属するか識別することができる

といった特徴があることを確認できたため，人間が自分では意識して学習をしなくても獲得することのできる知覚能力を検証する手がかりになると考える．

3. 実験の手順

破裂子音の有声・無声のミニマルペア（例えば/d/と/t/）は，口唇の開放から声帯振動までのタイミングの相対的な違いによってカテゴリカルに知覚されることが分かっている[4] . このタイミングを，「有声開始時間（Voice Onset Time ,VOT）」という．また，聴覚心理学の研究により，乳児も VOT を手がかりとした子音の聞き分けができ，さらに乳児にも子音カテゴリカル知覚の特徴がみられることが分かっている[4].

そこで本実験では，VOT 値を SOM に入力し，自己組織化によって有声・無声子音の識別境界が形成される様子を検証した．

3.1 実験条件

VOT は時間の値なので離散値となり，そのまま SOM に入力した場合，抽出できる特徴量が出現頻度しかなく，入力値同士の相関関係が学習されない．この問題を解決するために，VOT 値の近傍の値をセットにして入力することを考えた．例えば，入力 VOT 値が+20ms のとき，+17ms，+18ms，+19ms，+20ms，+21ms，+22ms，+23ms にして入力するようにした．

この入力値の近傍は，小さすぎると学習時にカテゴリが形成されにくくなり，大きすぎるとカテゴリの境界が曖昧になる．本実験では，後述する学習実験の回数をもとに近傍の範囲を決定しているが，人間の聴覚神経において，VOT がどのように符号化されているのかを厳密にモデル化したわけではないので，これについては議論の余地があると考えている．

また，SOM はカテゴリの自己組織化において入力値の出現頻度分布に大きく影響を受けるため，SOM を用いた検証実験を行う際に入力値の頻度分布の設定が問題となる．この問題を考慮して，以下の 3 つの実験条件を設定した．

(1) 条件 1：アメリカ人話者データに基づく実験

図 1 は，正常なアメリカ人英語話者 1 名が/t/と/d/で始まる複数の語を繰り返し発話したときの VOT 値の頻度分布である[4]. VOT の発生頻度は均等ではなく，偏りが見られる．頻度数の多い VOT 値は代表的な値と考えられる．

* 宮澤 幸希，早稲田大学大学院人間科学研究科修士課程，
埼玉県所沢市三ヶ島 2-579-15 e-mail：m-kouki@moegi.waseda.jp

図 1 の出現頻度をもとに自動生成した VOT 値を 1940 回 SOM に入力して学習を行った.

この学習回数は, 強化学習されたチンチラ(ネズミの一種)が人間と同様の子音カテゴリカル知覚を獲得できることを示した Kuhl＆Miller の論文[5]において, チンチラに与えた学習刺激となる自然音声の数をもとにしている.

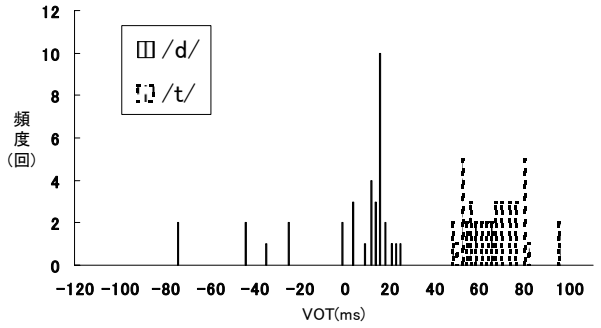


図 1.アメリカ英語話者 1 名の VOT 値頻度分布
(文献[4],p43 の図を書き直したもの, 原論文は Blumstein, et al, "Production Deficits in Aphasia : A Voice Onset-Time Analysis";1980)

(2) 条件 2:話者別・母音数別の頻度を考慮して, 動物実験と同条件で実験

論文[5]の実験を追試し, 実データに基づく強化学習の過程を検討するため, この実験と同じ手順に従って 1940 回の学習を行った. 論文[5]には, チンチラに学習させた VOT の値や種類, 話者の数, 学習の手順まで詳細に記録されている. また, 獲得されたカテゴリの境界値も示されており, 参考とした.

条件 2 の VOT 頻度分布を図 2 に示す. 論文[5]によると, チンチラの学習刺激となる音声のうち, カテゴリ/d/に相当する VOT 値は-200ms～+24ms, カテゴリ/t/に相当する VOT 値は +40ms～+128ms の範囲に分布していた. そこで, この範囲内で各話者の VOT 発話頻度は平均値の異なる正規分布に従うと仮定して, 話者と母音の種類によって VOT の出現頻度を変化させることで, 論文[5]の実験条件を再現した.

なお, 話者ごとの VOT 発話頻度平均値は, 各カテゴリの VOT 値の範囲内で均等に分散するとし, 簡易的な正規分布に従う形で頻度分布を作成した.

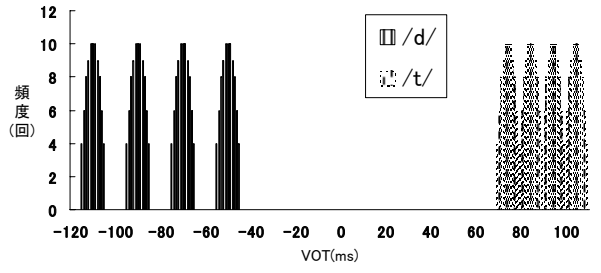


図 2. 実験条件 2 の VOT 値頻度分布
(論文[5]をもとに作成, 話者 4 人, 母音 3 種類のとき)

(3) 実験条件 3:条件 1,2 を混合した実験

条件 1, 2 をもとに, 図 3 に示した VOT の出現頻度モデルを作成して, 1940 回の学習を行った.

条件 1 は実際の発話に基づいているため入力値の信頼性は高いと考えられるが, チンチラの実験の設定条件とは異なるので, チンチラの獲得した知覚境界との比較が難しい. 一方, 条件 2 はチンチラの実験条件に従う形で学習を進めるので学習結果を評価しやすいが, 論文[5]にはチンチラに与えた VOT 値の範囲は示されているものの出現頻度は不明であるため, VOT 値の出現頻度は予測に過ぎない.

そこで, 条件 3 では, 1, 2 の実験条件を両方取り込み, 図 1 に示した一人の話者による VOT 値の頻度分布と, 論文[5]に示された複数の話者による VOT 値の範囲を組み合わせることで, VOT 入力値の範囲・頻度の両方をより実際の言語環境に近づけることを考えた. 条件 3 の VOT 頻度分布を図 3 に示す.

例えば, 図1では/t/カテゴリの VOT= +50msec 付近に頻度の山があるが, 論文[5]では VOT= +40ms～+64ms の範囲は全 12 回の試行中 1 回しか発音されておらず, 出現頻度が低いと考えられるため, 図 3 の頻度モデルでは/t/カテゴリはまとめて一つの山とした. 逆に/d/カテゴリの VOT 値は図 1, 論文[5]ともに広範囲にわたっているので, 図 3 のモデルでも広範囲に分布するようにした.

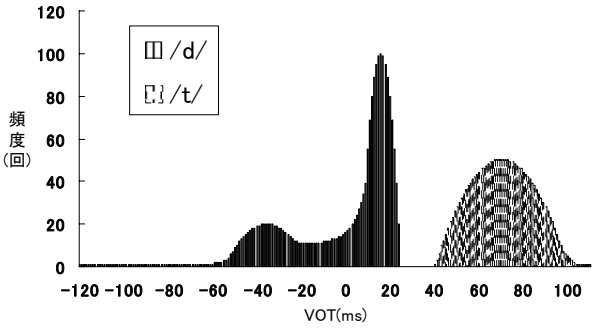


図 3. 実験条件 3 の頻度モデル

3.2 評価方法

学習を終了した SOM に, /d//t/のカテゴリカル知覚の重要な手がかりになっていると思われるいくつかの VOT 値を入力してカテゴリ形成の様子をみる. この VOT 値は, 条件 3 の頻度モデルの極大値・極小値, 条件 2 の入力値を参考に, 9 点とした(これらを指標 VOT と呼ぶ). 値は,

- D1 : VOT = -88ms (/d/カテゴリ)
- D2 : VOT = -35ms (/d/カテゴリ)
- D3 : VOT = -10ms (/d/カテゴリ)
- D4 : VOT = +16ms (/d/カテゴリ)
- D5 : VOT = +24ms (/d/カテゴリ)
- T1 : VOT = +40ms (/t/カテゴリ)
- T2 : VOT = +64ms (/t/カテゴリ)
- T3 : VOT = +76ms (/t/カテゴリ)
- T3 : VOT = +95ms (/t/カテゴリ)

とした.

SOM に VOT= -200ms～+127ms までの VOT 値を順番に提示していき, 各値が指標 VOT の/d//t/カテゴリのうち, どれに属しているかを調べることで, /d//t/カテゴリの境界となる VOT 値を決定する.

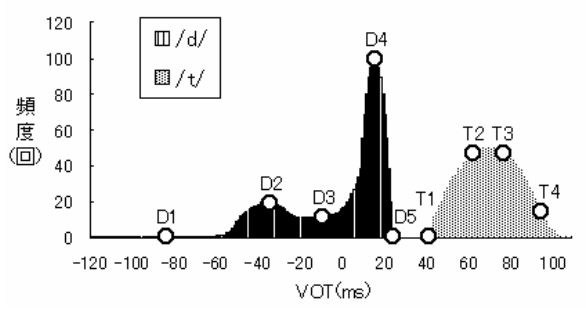


図 4. 指標 VOT

4. 実験結果

学習条件 1～3 について，それぞれ学習を 5 回ずつ行い，学習後のカテゴリ境界の値を表 1 に示した．先行研究[5]では，チンチラのカテゴリ境界は+33.5ms，人間のカテゴリ境界は+35.2ms が得られている．また，乳児では，VOT= +20ms から+40ms の間にカテゴリの境界があったことが示されている[4].

表 1 に示したように，カテゴリ境界の平均値は，3 条件全て乳児の子音カテゴリカル知覚実験の境界値の範囲内に収まった．さらに，チンチラ，人間のカテゴリ境界にも近い値となった．

表 1.各条件のカテゴリ境界値

	1回目	2回目	3回目	4回目	5回目	平均	分散
条件 1	31.5	29.5	41.5	31.5	58.5	38.5	3.29
条件 2	-41.5	61.5	61.5	-81.5	89.5	34.2	7.26
条件 3	33.5	30.5	34.5	35.5	30.5	32.9	1.43

(ms)

図 5 は，最も適当な学習の結果を示したと考えられる条件 3 (2 回目)の知覚境界をグラフにしたものである．図中の Human, Chinchilla は，論文[5]に示された人間の/d//t/知覚境界と，チンチラが学習によって獲得した/d//t/知覚境界である．

横軸は入力 VOT，縦軸は入力された VOT を/d/と同定した割合である．この値が 100%なら/d/カテゴリ，0%なら/t/カテゴリとして知覚されたことになる．

図 5 では，/d/と/t/の知覚カテゴリは VOT= +30ms～+40ms の区間で急激に変化しており，この区間に知覚境界があるといえる．すなわち，VOT= +30ms～+40ms の間で音刺激に最も大きな差異があるように感じられたということであり，カテゴリカル知覚の特徴が示されている．

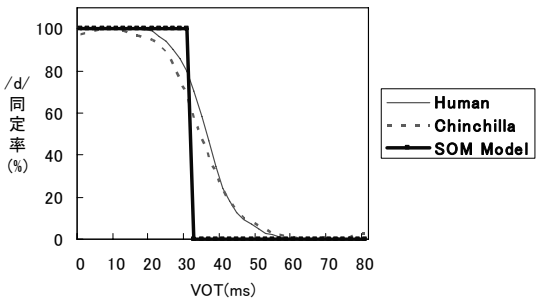


図 5.学習条件 3 の知覚境界

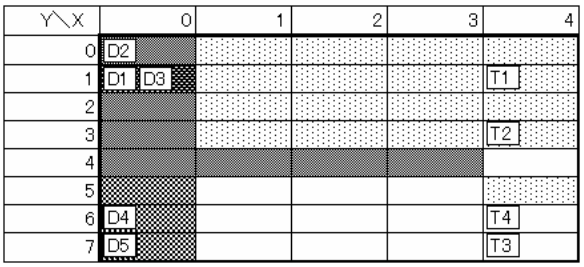
SOM の学習結果を，図 6 に示す．

(a)において， X 軸方向で/d//t/カテゴリが弁別されており，X 座標 0 付近が/d/，X 座標 4 付近が/t/に対応する．また，Y 軸方向で入力値の出現頻度が弁別されており，出現頻度が少ない値は Y 座標 0 付近，頻度が多い値(図 3 の頻度モデルの山の部分とその周辺)は Y 座標 7 付近に集まっている．

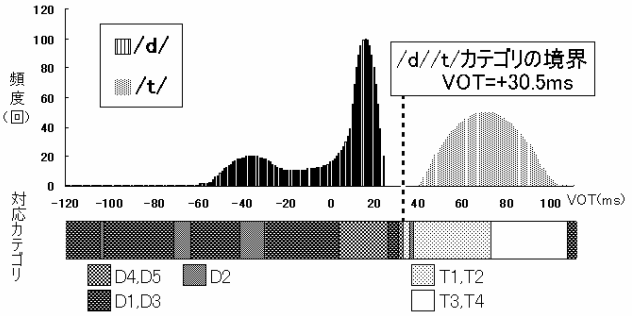
VOT 値の出現頻度モデルと，各値に対応するカテゴリを(b)に示す．図 3 の頻度表の2つの山と3つの谷がそれぞれカテゴリに分けられていて，

- VOT = -120～+2ms (カテゴリ D1, D2, D3)が入力頻度の少ない/d/
- VOT = +3～+30ms (カテゴリ D4, D5)が入力頻度の多い/d/
- VOT = +31～+38ms (D, T カテゴリ混在) は各カテゴリが混ざり合う/d//t/カテゴリの境界領域
- VOT = +39～+100ms (カテゴリ T1, T2, T3, T4)が入力頻度の多い/t/

となっており，ほぼ正しい/d//t/カテゴリの弁別ができている．



(a) SOM の出力層



(b) 各 VOT 値の出現頻度と，それぞれの値に対応するカテゴリ

図 6.条件 3 の学習結果

5. まとめ

学習を終えた SOM は，連続した入力値を/d/カテゴリ，/t/カテゴリに分離した．さらに，SOM による学習で人間に近い有声音・無声音の知覚境界を獲得できることが実証された．

本実験の結果だけでは，人間が生得的なメカニズムなしで，学習のみによって子音の知覚能力を獲得できるということは主張できないが，特殊な言語処理機構をもたない神経回路モデルと学習データのみを用いた実験で子音カテゴリカル知覚の再現ができたことから，言語獲得における自己組織的なメカニズムと言語経験の重要性を示すことができたといえる．

なお，本研究の学習実験では，必要なパラメータである VOT を初めから設定してやり，SOM は入力情報から VOT の境界値を学習するのみであった．

有声・無声子音の弁別に VOT が手がかりとなるという知識を人間の乳児が生得的に持っているのか，学習によって獲得した

のかについては議論の余地があるが，今後破裂子音全体，子音全体，母音を含めた音節，と弁別学習の対象を広げていくにあたって，学習に必要なパラメータは増えていくと考えられる．そのため，パラメータを自己組織的に選択できる学習モデルを開発する必要がある．

また，「ある環境で学習したデータをどのように新環境に適応させるか」についても検討を行い，学習モデルと環境適応アルゴリズムを統合した新しい音声言語理解技術の開発を目指したいと考えている．

参考文献

[1] 重野純: 音の世界の心理学，ナカニシヤ出版，2003.
[2] Gloria J. Borden, Katherine S. Harris: ことばの科学入門，メディカルリサーチセンター，1994.
[3] T・コホネン: 自己組織化マップ，シュプリンガー・フェアラーク東京，2005.
[4] ジャック・ライアルズ: 音声知覚の基礎，海文堂，2003.
[5] Kuhl, P. K. and Miller, J. D.: Speech Perception by the Chinchilla: Voiced-Voiceless Distinction in Alveolar Plosive Consonants，Science，1975.
[6] M・シュピッツァー: 脳　回路網の中の精神，新曜社，2001.