1D2-1

連続音声からの音韻カテゴリ獲得モデルに関する考察

A Consideration of Self-organized Phoneme Acquisition Model from Continuous Speech

宮澤幸希[*1*2] MIYAZAWA Kouki　三浦英朗[*1] MIURA Hideaki　菊池英明[*1] KIKUCHI Hideaki　馬塚れい子[*2] MAZUKA Reiko

[*1] 早稲田大学 人間科学学術院 Faculty of Human Sciences, Waseda University
[*2] 理化学研究所 脳科学総合研究センター RIKEN Brain Science Institute

It is unclear as to how infants learn the acoustic expression of each phoneme of their native languages. We use a natural continuous speech and build a self-organization model that simulates the cognitive ability of the humans, and we analyze the quality and quantity of the speech information that is necessary for the acquisition of the native phoneme system. Our model is designed to learn values of the acoustic features of a continuous speech and to estimate the number and boundaries of the phoneme categories without using explicit instructions. In a recent study, our model could acquire the detailed vowels of the input language. In this study, we examined the mechanism necessary for an infant to acquire all the phonemes of a language. As a result, we showed that the acquisition of an unstable phoneme was possible without the use of instructions.

1. はじめに

健常なヒトは，ある言語環境で生活しているだけで，誰でも母国語の音韻を聞き分けることができるようになる．ヒトの言語獲得は，対象となる言語の詳細な情報（対象言語の音韻の数と種類）を必要としない教師なし学習であるといえる．我々の目的は，このようなヒトの優れた音韻獲得能力のメカニズムを解明し，音声言語処理技術に応用することである．先行研究では，計算論モデルによる音韻獲得の検証が行われてきたが，これらの研究では限定された語彙の音声を入力としていた．

そこで我々は，自然な連続音声の入力と，ヒトの認知能力を模擬した自己組織化学習モデルを使って，音韻体系の獲得に必要な音声情報の種類と量に関して検討している．これまでに，本モデルを使って単独話者の日本語音声から，日本語固有の母音体系を獲得可能であることを示した[Miyazawa 2010]が，このモデルは学習において入力の音響的な定常性の影響が大きく，母音の学習に適していた．そこで本研究では，全音韻を獲得可能なモデルのメカニズムに関して検討する．音響的な定常性が異なる母音と子音を取り扱うため，ヒトの聴覚系の処理に基づく動的特徴フィルタを導入したモデルを提案する．

次章で先行研究を紹介し，本研究の試みについて述べる．3章では我々のモデルの詳細について述べる．4 章では提案モデルによる音韻獲得実験の詳細を，5 章ではまとめを述べる．

2. 先行研究

2.1 母国語の母音体系の獲得

ヒトは生後 6ヶ月以内に母国語の母音の弁別能力を獲得し，12 ヶ月以内に子音の弁別能力を獲得する．短期間の学習においても，乳児は/ta/から/da/まで連続的に変化する音声刺激の出現頻度分布を学習できることが分かっており，分布が二峰性の刺激セットを 2 分間聴取した乳児のグループのみ，/ta/及び/da/を弁別することができた[Maye 2002]．これらの知見から，乳児は成人の音声の特定の音響特徴の統計的分布を利用して母国語の音韻境界の学習を行っているという仮説が提案された．いっぽうで，言語獲得は生得的な機構によるという仮説もある．

連絡先：宮澤幸希，早稲田大学人間科学学術院，〒359-1192 埼玉県所沢市三ヶ島 2-579-15，048-462-1111 (ext:6758)，m-kouki@moegi.waseda.jp

2.2 音韻学習の計算論的モデル

この議論に対するアプローチの一つとして，ヒトの言語獲得をモデル化し，音声情報の何をどの程度使えば，ヒトの知覚能力を再現出来るのかを検証する試みがある．例えば，F2,F3 を特徴量とした競合ヘッブ学習による/r/, /l/の弁別[Guenther 1996]，F1,F2 及び持続時間を特徴量としたガウス混合分布モデルによる長短母音の弁別[Vallabha 2007]などがある．これらは，教師なし学習によって言語固有の音韻の獲得に成功しており，言語獲得における言語経験の重要性を示している．しかし先行研究では，音韻の出現頻度が統制された刺激セットと，特定の音韻の弁別のために事前に解析された特徴量を使っている．そのため，乳児の言語環境に比べて学習が容易である可能性がある．

そこで我々は，より自然な入力と，ヒトの聴覚表現に近い特徴量および学習モデルを使って，言語固有の母音体系の数と境界の獲得過程を検証してきた．我々のモデルは明示的な教示無しで音韻カテゴリーの数や種類を推定することができる．本モデルによるシミュレーションの結果，日本人話者の 100 秒の自然な連続音声からヒトと同等の母音体系を獲得することができた．我々のモデルの学習結果を表 1 に，ヒトの切り出し母音に対する母音同定率[桑原 1972]を表 2 に示す．表の in が提示した母音，out が認識された音韻を示す．others は母音以外に分類された数を示す．/e/,/o/,/a/の認識率が高いなど，我々のモデルはヒトの知覚と類似した結果を示した．これは自然な音声は乳児が利用可能な形で，母音体系の情報を含むことを示唆している．

表 1. 我々のモデルの母音同定率(%) [Miyazawa 2010]

in＼out	i	e	o	a	u	others
i	56.1	23.7	4.5	0.2	6.1	9.4
e	9.2	79.0	0.4	0.8	4.2	6.4
o	0.9	1.6	66.1	8.5	5.4	17.5
a	0.3	0.9	6.7	71.9	2.0	18.2
u	9.7	14.8	4.8	2.7	41.2	26.8

表 2. 成人の母音同定率(%) [桑原 1972]

in＼out	i	e	o	a	u	others
i	52.0	0.0	0.0	0.0	2.0	46.0
e	4.0	70.0	3.0	12.0	5.0	6.0
o	0.0	1.0	58.0	11.0	25.0	3.0
a	0.0	1.0	0.0	57.0	2.0	40.0
u	10.0	3.0	1.0	1.0	53.0	32.0

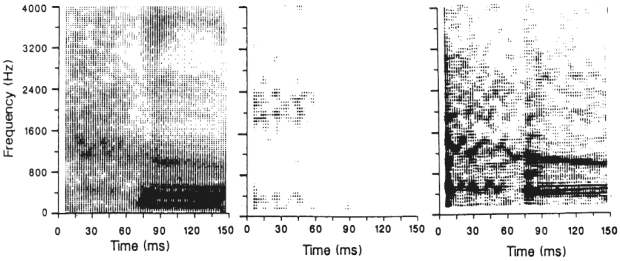

図 1: 音声/pu/に対する応答 [Carney 1986]

ただし, [Miyazawa 2010]のモデルは母音及び定常的な特徴を持つ子音(鼻音など)は独立したカテゴリーを形成したものの, 破裂音などの非定常な子音は獲得できなかった. [Miyazawa 2010]のモデルは入力の頻度と分布に基づいてクラスタリングを行うが, 非定常な子音は短期的なスペクトルの変動が多く, 出現回数も少ないため, 独立したクラスタを形成しにくい. この問題を解決するために, 本研究では学習の前処理として, ヒトの聴覚抹消系における非定常な音響特徴の処理過程を近似する.

2.3 聴覚系における動的特徴抽出とそのモデル

ヒトの末梢聴覚細胞の応答頻度に基づいて音声の動的な音響特徴の抽出が行われている可能性が指摘される. 図 1 は, 音声/pu/に対する一次聴覚神経の応答である. 左図は入力のスペクトログラム, 中図はスペクトルの変動の大きい子音部で応答を示す「高頻度細胞群」, 右図はスペクトルが定常な母音部などで応答を示す「低頻度細胞群」を示す[Carney 1986]. また, 抹消聴覚系において, 数十 ms 単位で音素の処理, 百 ms 単位で音素列の処理など, 別々の経路で異なるタイムスケールの処理が行われている可能性がある[Hickok 2007]. これらの知見に基づき, 本研究では前処理としてスペクトルの変動の大きさに基づいて音声入力の分離を行うことで, 各音韻の出現頻度が異なるために統一のモデルで扱えない問題を解決する. 具体的には, 3.2 で述べる⊿MFCC が聴覚系における動的特徴抽出の近似として妥当であると仮定した. 類似手法として, パワーを利用して連続音声から母音の分離を試みる研究[高良　2009]が一定の成功をおさめており, 本手法の妥当性が支持される.

3. 連続音声からの音韻獲得モデル

3.1 教師なし学習のアルゴリズム

教師なし分類のアルゴリズムとして, 自己組織化マップ(SOM)を使用する. SOM は大脳感覚野における感覚特徴表現が知覚経験によって形成される過程を再現したニューラルネットワークの一種である[Kohonen 1990]. SOM は多次元の入力を教師無しでクラスタリングし, カテゴリーを推定するため, 音韻の獲得モデルとして適当であると考える. また, 一般的な SOM は入力をノード数と同数のクラスに分類するが, 乳児は母音の数も学習している. そこで, 類似したカテゴリーを同一カテゴリーに統合する枠組みを導入した. ユニット(ノード)が 1 列に配置された一次元(直線形)の SOM については, 学習結果の密度ヒストグラムによってカテゴリーを推定する手法[寺島 1996]が提案されており, これに基づいてカテゴリー数を評価した. 本モデルのアルゴリズムは[Miyazawa 2010]で詳述している.

3.2 前処理

原音声を, 従来の音声処理技術で広く利用される音響特徴量であるメル周波数ケプストラム係数(MFCC)に変換する.

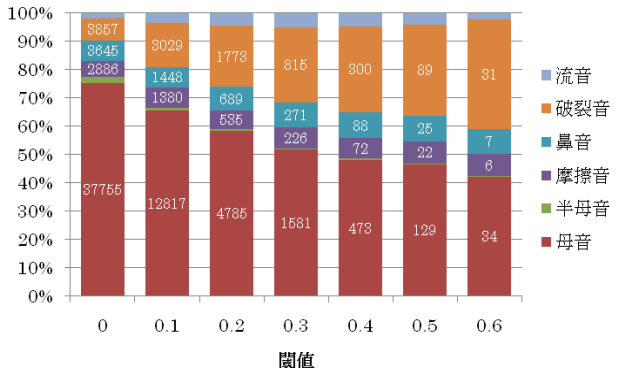

図 2: 全フレームに対する各音素の割合

MFCC は聴覚系の処理に基づき, 全音韻の認識に十分な情報を含む. 本研究では, MFCC 12 次元に対数パワー(C0), ⊿MFCC 12 次元, ⊿C0 を加えた計 26 次元(MFCC26)を使用し, フレーム長 25ms, フレームシフト長 10ms で解析を行っている.

学習の前処理として, 2.3 で述べたように動的特徴に基づき定常・非定常音声に分離する手法を提案する. 工学的に定常な音声と非定常な音声の分離を行う方法として, ⊿MFCC の二乗和をとる方法[Hodoshima 2006]がある. ⊿MFCC は音声の動的な変化を表現するために提案された特徴量で, 連続した MFCC の回帰係数である. d を⊿ケプストラム, t をフレームサイズ, Θ をフレームの総数, C を MFCC とすると,

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta\left(C_{t+\theta} - C_{t-\theta}\right)}{2\sum_{\theta=1}^{\Theta} \theta^2} \qquad (1)$$

本研究では HTK にならい, t=2 とした. 本研究では, ⊿MFCC の 1～12 次の二乗和の値を求め, 特定の閾値以下であれば MFCC の変化が少なく定常, 閾値以上であれば非定常であると判断する. この処理を「動的特徴フィルタ」と呼ぶ.

動的特徴フィルタの適切な閾値を得るために, 日本語話し言葉コーパス(CSJ)[Maekawa 2003]を使用して予備的な評価を行った. 男女 2 名ずつ, 計 4 つの音声の MFCC26 を求め, 各フレームを動的特徴フィルタにかける. 結果を図 2 に示す. 各バーの内訳は動的特徴フィルタによって取り出された全フレームに対する, 各音韻のフレームの割合と数を示す. 横軸は閾値, 縦軸は割合である. グラフ中の数値はフレーム数を示す. 閾値が増えるに従って, 母音や鼻音などの定常な音韻は全体に占める割合が減少するのに対し, 破裂音などの非定常な音韻は割合が増加している. したがって, 動的特徴フィルタによって非定常な音韻を抽出できたといえる. 本研究では, 動的特徴フィルタの適切な閾値の値を 0.2 と仮定して次章の実験を行った.

4. シミュレーション

4.1 手続

乳児が言語獲得の際に最初に親の声を聞くと仮定して, 単独話者が自然な環境で発話した連続音声を入力とする. 事前に動的特徴フィルタによって定常音声, 非定常音声に分離し, それぞれを異なる SOM モデルに入力し, 音韻の数と境界が形成される過程をシミュレーションする. 以下では定常音声(閾値以下)を入力するモデルを定常 SOM, 非定常音声(閾値以上)を入力するモデルを非定常 SOM と呼ぶ.

表 3. 学習に使用した音韻カテゴリーの詳細

Group	Phoneme	Group	Phoneme
vowel	/i/,/e/,/o/,/a/,/u/	voiceless plosive	/k/,/t/
semivowel	/y/	voiced plosive	/g/,/d/
fricative	/s/	liquid	/r/
Nasal	/m/,/n/		

CSJ より, 20～30 代の男女 5 名ずつ, 計 10 名の話者を選んで入力音声とし, 3.2 にならって解析した. [Miyazawa 2010]において, 入力音声のスピーチスタイルが異なっていても学習結果に有意な差はなかったので, スピーチスタイルは Simulated Public Speaking に統一した. 各データの連続した 10320 フレーム(103.2 秒)を学習用データ, 残りを評価用データとして 5 回のクロスバリデーションの試行を行った. 訓練データを動的特徴フィルタにかけ, 定常 SOM と非定常 SOM の学習を行う. 両 SOM のノード数は 28 とし, 1 回の学習(ステップ)では, 訓練データ全体から順番に 1 フレームずつ入力する. データを 2 往復, 計 20640 ステップの学習が完了した時点で訓練を終える.

4.2 評価

以下の方法で SOM が獲得した音韻の数と境界を評価した. はじめに, 評価用データの各音素の中心 1 フレームを切り出して音韻ラベルをつける. 各音韻カテゴリーの音素数は最も数の少ない音韻に合わせた. 評価対象の音韻のリストを表 3 に示す.

続いて, 3.1 に基づいてクラスタを統合した定常 SOM と非定常 SOM に対して, 以下のようにして各音韻の認識率を求める.

(1) 各音韻の各クラスタにおける占有率(同じ音韻の評価データの何割がそのクラスタに対応付けられたか)を求める.
(2) 各クラスタについて最も占有率の高い音韻を求め, そのクラスタの音韻ラベルとする. 対応クラスタの無い音韻が存在した場合, 占有率の高い順にクラスタを検索し, どの音韻にも対応していないクラスタに対応付ける. 該当クラスタが存在しない場合, その音韻の認識率は 0%とする.
(3) 各クラスタが, 対応する音韻ラベルの評価データの何割と対応付けられているかを求め, その値が認識率となる.

集計の際には, 認識率が 0 になった試行は除外して各話者の音韻カテゴリーごとに平均認識率を求めた.

4.3 結果

図 3 に, 訓練終了後の認識率を示した. 各値は 10 名の話者の平均認識率である. 縦棒は標準偏差を示す. 各音韻グループの左のバーは非定常 SOM の結果, 中央のバーは閾値を導入しない従来のモデル[Miyazawa 2010]の結果, 右のバーは定常 SOM の結果を示す. 無声破裂音($F(2,27)=3.35$, $p<.001$), 有声破裂音($F(2,27)=3.35$, $p<.05$)において, SOM モデル間で有意に認識率の違いがみられた. その他の音韻グループではモデル間の有意差はなかった.

破裂音はこれまでのモデルでは取り扱いが困難であったが, 非定常 SOM の導入により学習可能であることが示された.

5. まとめ

ヒトが母国語の音韻体系を獲得する過程を解明することを目的として神経回路網モデルを使用して学習実験を行った. 我々のこれまでのモデル[Miyazawa 2010]では非定常な子音の学習が困難であったため, ヒトの聴神経系の機能を近似した動的特徴フィルタを導入し, 音声・非定常音声を異なるモデルに入力して学習を行った. 実験の結果, これまでのモデルでは学習が困難であった破裂音の同定率が 10～15%向上した.

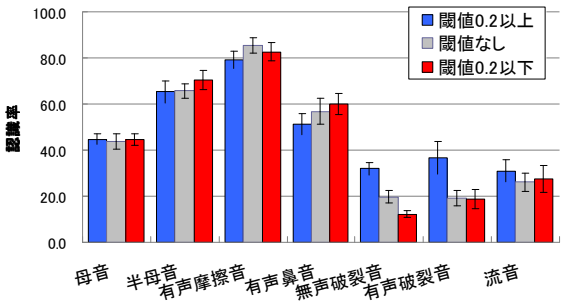

図 3: 動的特徴を導入したモデルの学習結果

本研究の結果から, 乳児が利用可能な音声資源及び認知機構によって, 約 100 秒の短い音声から言語固有の母音・子音体系の情報を抽出できる可能性が示された. これは, 乳児が成人の音声の統計的情報を学習しているとする仮説を支持する.

参考文献

[Miyazawa 2010] Miyazawa, K., Kikuchi, H., Mazuka, R. : Unsupervised Learning of Vowels from Continuous Speech based on Self-organized Phoneme Acquisition Model , INTERSPEECH2010, 2010.

[Maye 2002] Maye, J., Werker, J. F., and Gerken, L. : Infant sensitivity to distributional information can affect phonetic discrimination, Cognition., 2002.

[Guenther 1996] Guenther., F. H., and Gjaja., M. N. : The perceptual magnet effect as an emergent property of neural map formation, J. Acoust. Soc. Am., 1996.

[Vallabha 2007] Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., and Amano, S. : Unsupervised learning of vowel categories from infant-directed speech, Proceedings of the National Academy of Sciences, 2007.

[桑原 1972] 桑原尚夫, 境久雄: 連続音声中の切り出し母音および音節の音韻知覚, 日本音響学会誌, 1972.

[Carney 1986] Carney, L. H., and Geisler, C. D. : A temporal analysis of auditory-nerve fiber responses to spoken stop consonant-vowel syllables, J. Acoust. Soc. Am., 1986.

[Hickok 2007] Hickok, G., and Poeppel, D : The cortical organization of speech processing , Nature reviews neuroscience, 2007.

[高良 2009] 高良富夫, 藤田祐貴, 砂川泰毅, 大城武志, 祝三志郎: クラスタ化を基本とする音声言語獲得の機能モデル : 単語と母音音素の場合, 電子情報通信学会技術研究報告. SP, 音声, 2009.

[Kohonen 1990] Kohonen, T : The self-organizing map , Porceedings of the IEEE, 1990.

[寺島 1996] 寺島幹彦, 白谷文行, 山本公明: 自己組織化特徴マップ上のデータ密度ヒストグラムを用いた教師なしクラスタ分類法, 電子情報通信学会論文誌. D-II, 1996.

[Hodoshima 2006] Hodoshima, N., Arai, T., Kusumoto, A., and Kinoshita, K. : Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments, J. Acoust. Soc. Am., 2006.

[Maekawa 2003] Maekawa, K. : Corpus of Spontaneous Japanese : Its Desugn and Evaluation, Proceedings of ISCA and IEEE SSPR, 2003.