音素の合成性を仮定した連続信号をサインとした分散的ベイズ推論に基づく記号創発

齋藤一誠 [1]　劉智優 [1]　長野匡隼 [1]　中村友昭 [1]　谷口彰 [2]　谷口忠大 [3]
[1] 電気通信大学 [2] 立命館大学 [3] 京都大学
i_saito@radish.ee.uec.ac.jp

概要

エージェント間のコミュニケーションを通して共有されるサインを創り出すことができる記号創発モデルが提案されている．これまでの研究ではサインの生成・認識能力の学習過程はモデル化されていなかった．そこで我々はこれまでに連続信号をサインとして用いた記号創発モデルを提案した．その手法では，複数の音素を組み合わせて単語を構成する音素の合成性を導入した連続信号から，エージェント間で対象を表現するサインを学習する．本稿ではその手法の比較・評価を行った．実験では，各エージェントは同一の物体を観測し，その物体を表現する物体の潜在表現と，それを表現するサインである連続信号を，コミュニケーションによって推論する．提案手法がエージェント間で高い一致率かつ高い精度で物体を表現するサインを学習可能であることを示した．

1　はじめに

人間が他者や環境と相互作用する中で，意図伝達可能な記号が創発する過程を創発コミュニケーションと呼ぶ．また，特に人の言語創発では，二重分節性（duality of patterning）と呼ばれる構造的性質を有するサインが用いられている [1]．二重分節性では意味を持たない音声の最小単位（音素）が組み合わさることで有意味な単語を形成し，それらの単語が組み合わさることで，より複雑な意味表現が可能な文を形成することができる [2, 3]．創発コミュニケーションの研究においては，有意味な単語を組み合わせて文を形成する研究が数多く行われている．一方，無意味な音素から有意味な単語を形成する研究はきわめて少ない．そのため創発コミュニケーションにおける音素から単語を形成する意義や役割については，未だ十分に解明されていない．そこで，本研究はこの音素から単語を形成する過程に焦点を当て，構成論的なアプローチを用いて音素を組み合わせて単語を創り出す音素の合成性の役割を解明する一助になることを目指す．

人間は，形成した単語や文を音声という連続信号を介してコミュニケーションする．そのため，話し手が同じ単語を発声したとしても，話し手自身の発声のばらつきやノイズの影響で，受け手が聞き取る音声信号は完全には一致しない．このような状況において，合成性がない場合，様々な事象を区別するためには，膨大な数の異なる連続信号が必要となり，その生成・認識が困難となると考えられる．一方で，合成性がある場合は，限られた種類の音素を組み合わせることで膨大な種類の音声信号を作り出すことができる．すなわち，限られた種類の音素の生成・認識ができればよく，音声言語に含まれる揺らぎに対して頑健な生成・認識が可能となると考えられる．本稿では，これを仮説として構成論的に検証することを目的とする．

この仮説を検証するためには，環境と音素列から生成した連続信号を相互に分節化することで，記号が創発する過程を再現可能なモデルが必要である．これまでの記号創発の研究として，谷口は集合的予測符号化仮説を提案し [4]，他者の内部状態を直接観察することなく，独立した 2 者の観測を表現する記号を推論する方法として，メトロポリス・ヘイスティングス名付けゲーム（MHNG）法を提案した [5]．しかし，コミュニケーションで使用されるサインは離散変数で表され，連続信号を用いた記号創発をモデル化できていなかった．人は自身の内部状態を他者へ伝えるための信号を作るサインを生成する能力と，それを解釈して自身の内部状態と結びつけるサインを認識する能力を学習する．つまりこの設定ではサインの生成および認識する能力が事前に備わっており，学習の初期段階から常にサインを誤り

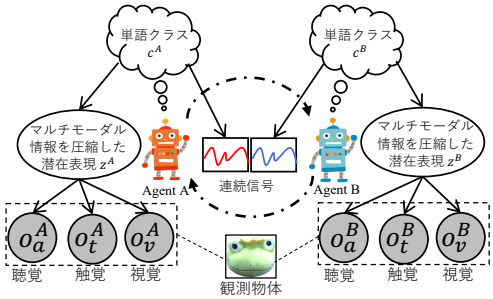

図 1　提案手法の概要

なく他者へ伝達できる前提になっていた．

一方人間は，記号創発の過程で連続信号（音声）を分節化することで，そのようなサインの生成・認識能力も学習している．生後すぐには音声の生成・認識をすることはできないが，他者とのインタラクションを通じて連続した音声を分節化し，音素や単語というパターンを見つけ出す．このような能力は，連続信号から頻出するパターンを抽出することが可能な Nonparametric Bayesian Double Articulation Analyzer (NPB-DAA) や Gaussian process hidden semi-Markov model (GP-HSMM) に代表される教師なし分節化モデル [6, 7, 8, 9, 10, 11] を利用することで再現可能である．しかし，これらの研究では，既に確立された音声言語が存在することを前提として，教示者の音声信号からその言語を構成している音素と単語を学習する過程の再現に留まっている．そこで本稿では，記号の創発過程をモデル化できる MHNG と，連続音声を分節化し音素を抽出可能な GP-HSMM [7] を組み合わせ，サインを連続信号とした記号創発が可能な確率的生成モデルを提案する．

図 1 が提案手法の概要である．各エージェントは同一の物体を観測し，その物体を表現する潜在表現・単語クラス・連続信号を推論する．この推論は連続信号を介したコミュニケーションによって行われる．実験では，観測としてロボットが物体から取得した視覚，触覚，聴覚情報を使用し，単語クラスと連続信号，さらに連続信号を構成している音素が学習可能であることを示す．また音素の合成性を仮定した提案手法は，1 音素のみで観測を表現する手法に比べて学習精度が向上することを示す．

2　提案手法

2.1　生成過程

提案手法のグラフィカルモデルは図 2 である．このモデルでは，観測である聴覚・触覚・視覚情報 o_a,o_t,o_v と連続信号 S が，以下のプロセスによって

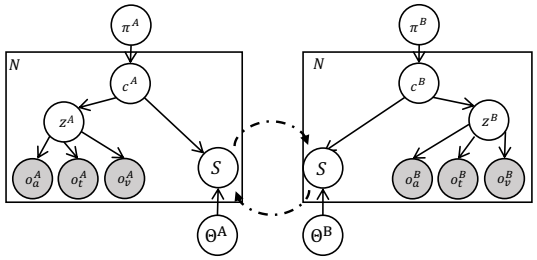

図 2　提案手法のグラフィカルモデル

生成されることを仮定している．まず π^A をパラメータとする多項分布から，単語クラス c^A を生成する．

$$c^A \sim p(c^A|\pi^A). \tag{1}$$

この単語クラスは，物体を表現する潜在表現 z^A と，その物体を表現する連続信号 S を抽象化した，エージェントの内部表現である．次に単語クラス c^A と対応する観測情報の潜在表現 z^A を生成する．

$$z^A \sim p(z^A|c^A). \tag{2}$$

潜在表現 z^A から観測 o_a^A,o_t^A,o_v^A を生成する．

$$o_a^A, o_t^A, o_v^A \sim p(o_a^A, o_t^A, o_v^A|z^A). \tag{3}$$

この生成モデルとして，Multimodal Latent Dirichlet Allocaiton [12] を用いる．また単語クラス c^A に対応する連続信号 S を，パラメータ Θ^A の GP-HSMM により生成する．

$$S \sim p(S|c^A, \Theta^A). \tag{4}$$

エージェント B もエージェント A と同様の生成過程である．

2.2　推論

MHNG に基づき，話し手によるサインの伝達，聞き手による採択率に基づくサインの受理・棄却によって提案モデルのパラメータを推論する．本章では，エージェント A を話し手，エージェント B を聞き手とした場合で説明する．図 3 が推論の概要である．

2.2.1　話し手による連続信号の生成

まず，エージェント A は，観測 o_a^A, o_t^A, o_v^A の潜在表現 z_A を推論する．

$$z^A \sim p(z^A|o_a^A, o_t^A, o_v^A) \tag{5}$$

この推論には MLDA[12] を用いる．次に，潜在表現を表す単語クラス c^A をサンプリングする．

$$c^A \sim p(c^A|\pi^A, z^A) \tag{6}$$

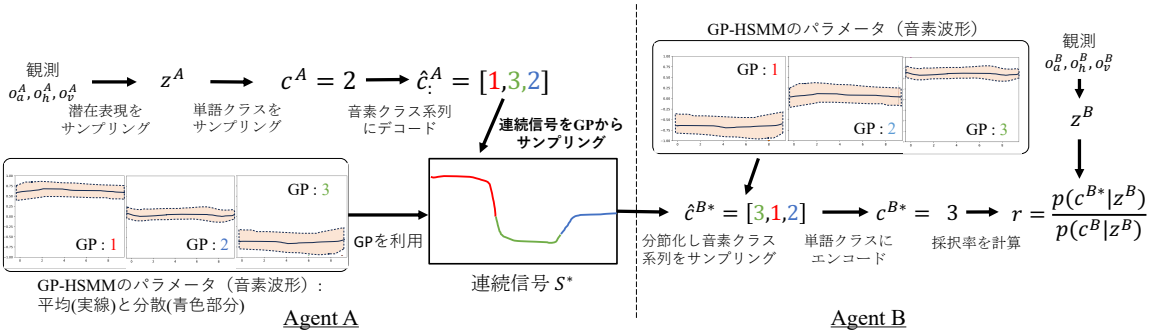

図 3　提案手法の推論の概要

連続信号は合成性を持ち，音素を組み合わせて生成されると仮定し，エージェント A の単語クラス c^A を L 個の音素クラス系列にデコードする．

$$(\hat{c}_1^A, \hat{c}_2^A, \cdots \hat{c}_L^A) \leftarrow \text{decoder}(c^A) \tag{7}$$

音素クラス系列から GP-HSMM で生成された音素 $s_1^A, s_2^A, \cdots s_L^A$ を接続することで，連続信号 S が生成される．L は文字数（音素数）を表しており，両エージェントで共有されることを仮定している．

$$s_l^A \sim \mathcal{GP}(s|\Theta^A, \hat{c}_l^A) \quad : l = 1, \cdots, L \tag{8}$$

$$S = \text{concatenate}(s_1^A, \cdots, s_L^A) \tag{9}$$

式 (7) と式 (9) は決定論的な処理となるため，式 (7)〜式 (9) は単語クラス c^A から連続信号 S をサンプリングしているとみなすことができ，次式のように表現することができる．

$$S \sim p(S|c^A, \Theta^A) \tag{10}$$

すなわち，式 (6) と式 (10) の二段階のサンプリングによって，連続信号 S を生成していると考えることができる．生成された S はエージェント B に送信される．

2.2.2　聞き手による連続信号の認識

エージェント B は信号 S を受け取り，GP-HSMM のパラメータを推論し分節化することで，音素クラス系列 $\hat{c}_1^B, \hat{c}_2^B, \cdots, \hat{c}_L^B$ をサンプリングする．

$$\hat{c}_1^B, \hat{c}_2^B, \cdots, \hat{c}_L^B \sim p(c_1, c_2, \cdots, c_L|S, \Theta^B) \tag{11}$$

式 (7) のデコードとは逆の処理により，音素クラス系列を単語クラスへエンコードする．

$$c^B \leftarrow \text{encoder}(\hat{c}_1^B, \hat{c}_2^B, \cdots, \hat{c}_L^B) \tag{12}$$

前節の連続信号の生成と同様に，式 (12) は決定論的な処理であるため，式 (11) と式 (12) はまとめて，連続信号 S から単語クラス c^B をサンプリングしているとみなすことができる．

$$c^B \sim p(c^B|S, \Theta^B) \tag{13}$$

2.2.3　聞き手による連続信号の受理/棄却

Metropolis-Hastings (MH) 法による潜在変数の推論のため，提案分布と目標分布を定義し，採択率を計算する．

前節までの信号の生成と認識では，式 (6) と式 (10)，式 (13) の三段階のサンプリングをしていると考えることができる．つまり，次式の同時確率から 3 つの変数をサンプリングしていることになる．

$$c^A, c^B, S \sim p(c^A, c^B, S|z^A) \tag{14}$$

$$= p(c^A|z^A)p(S|c^A)p(c^B|S) \tag{15}$$

$$\propto p(c^A|z^A)p(S|c^A)p(S|c^B) \tag{16}$$

$$\equiv Q(c^A, c^B, S) \tag{17}$$

ただし，2 式目から 3 式目の変形では $p(S)$ に一様分布を仮定している．この分布 Q を MH 法における提案分布とする．

学習の目標は両者の観測を表す潜在表現から，それぞれの単語クラス c_A，c_B と，共有されるサイン S を推論することである．すなわち，次式の事後分布からのサンプルを生成することが目的であり，これを目標分布とする．

$$p(c^A, c^B, S) = p(c^A, c^B, S|z^A, z^B) \tag{18}$$

$$= p(S|c^A, c^B)p(c^A|z^A)p(c^B|z^B) \tag{19}$$

$$\approx C \cdot p(S|c^A)p(c^A|z^A)p(S|c^B)p(c^B|z^B) \tag{20}$$

ただし，C は正規化定数であり，2 式目から 3 式目への変形では Product of Experts 近似を用いた．

式 (17) の提案分布と式 (20) の目標分布より，新たなサンプル c^{A*}, c^{B*}, S^* の採択率は次式となる．

$$r = \frac{Q(c^A, c^B, S)p(c^{A*}, c^{B*}, S^*)}{Q(c^{A*}, c^{B*}, S^*)p(c^A, c^B, S)} \tag{21}$$

$$= \frac{p(c^{B*}|z^B)}{p(c^B|z^B)} \tag{22}$$

すなわち，エージェント B のパラメータのみを利用した c^{B*} に関する評価のみで，サンプル c^{A*}, c^{B*}, S^* の受理と棄却を判断できることを意味している．

3　実験

両エージェントの観測として，文献 [13] で使用された，11 カテゴリ 67 個の物体からロボットが取得したマルチモーダル情報（視覚・聴覚・触覚情報）を用いた．連続信号の初期値として，−1〜1 の乱数によって生成した連続信号を各物体にランダムに割り当てた．学習は初期値を変え 10 回行なった．

サインの合成性を再現するため，3 種類の音素を組み合わせた音素系列長 $L = 3$ の連続信号を学習する提案手法（合成性有り）と，カテゴリ数と同数の 11 種類の音素を音素系列長 $L = 1$ の連続信号として学習する合成性を用いない手法 (合成性無し）を比較した．また性能上限としてサイン生成時に常に揺らぎのない信号を生成し，学習初期から誤りなくお互いのサインが伝わる設定 (性能上限）でも実験を行った．また MHNG を用いず，観測からサインを推論するエンコーダ（Sender）と，サインから観測を再構成するデコーダ（Receiver）を深層強化学習で学習する手法 (Auto-encoder game) をベースラインとした [14]．この手法では，Receiver の観測尤度を報酬として，Sender と Receiver を学習している．

3.1　実験結果

物体カテゴリを表現する単語クラスが学習できているか検証するため，各物体から推論された単語クラス c_A，c_B と正解の物体カテゴリ間の ARI を計算した．全 10 回の ARI の平均が表 1 である．合成性を仮定した連続信号のサインを用いた手法は，性能上限の誤りなくサインのやりとりが可能な手法と同等の精度になった．また音素の合成性を仮定した手法は合成性を仮定しなかった手法と比べて，0.2 ほど ARI が高かった．

Auto-encoder game は，Sender の学習に Receiver の尤度を利用できる設定となっている．他者の尤度を定量的に誤りなく利用できるこの設定は，独立した個体間のコミュニケーションの観点からすると，不自然な設定であるといえる．一方，提案手法は，他のエージェントから送られるサインと自己の内部状態のみを基に分散的に推論を行う，より難しい設定となっている．それにもかかわらず，提案手法の ARI は Auto-ecncorder game より高い値となった．

次に 2 体のエージェント間の単語クラスの一致度がどのように変化していくのか見るため，それぞれの手法での各エージェントの単語クラスの一致率を

表 1　単語クラスと正解物体カテゴリ間の ARI

	Agent A	Agent B
性能上限	0.81	0.81
合成性有り	**0.81**	**0.80**
合成性無し	0.60	0.60
Auto-encoder game	0.70	-

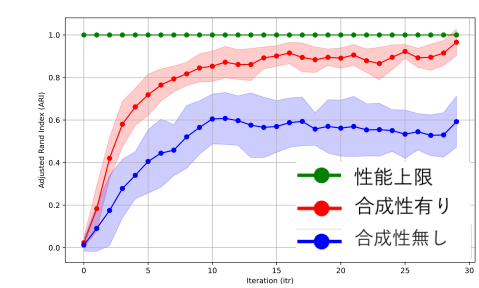

図 4　エージェント同士の単語クラスの一致度

ARI を用いて計算した．学習回数に対する一致率の変化が図 4 である．合成性を仮定した場合と仮定していない場合でグラフを比較すると，合成性を仮定した手法が分散も小さく安定してエージェント間の単語クラスを一致させることができている．このことより，2 体のエージェントは合成性を仮定したサインを用いたコミュニケーションにより，共通の単語クラスを作り出すことができたことが分かる．

これら 2 つの実験のいずれにおいても，合成性を仮定したサインを用いた提案手法は，性能上限と同等の精度を達成し，合成性を仮定しない手法と比較して大幅に精度が向上した．この結果から，生成されるサインが連続音声で話者内変動のような揺らぎを持つ場合，少ない種類の音素を組み合わせてサインを構成する提案手法が，多種類の音素のうち 1 音のみでサインを構成する手法よりも有効であることが示された．（音素の学習過程については参考情報 A に添付した．）

4　まとめと今後の課題

本稿では，連続信号をサインとして用いた記号創発モデルを提案し，エージェント間で共有概念とサインを学習する能力を検証した．実験結果から，音素の合成性を仮定した手法が，連続信号を用いた環境下でも高精度にエージェント間で共有される記号の創発が可能であることを示した．

一方で，提案手法は単語の長さが初めから固定されている不自然な設定になっている．よって今後は，今回の実験の考察を深めると共に，柔軟な単語長を創発するモデルへと拡張し，検証していく．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011 の支援を受けたものである．

参考文献

[1] André Martinet and Elisabeth Palmer. Elements of general linguistics. **Foundations of Language**, 2(2):151–186, 1966.

[2] W Tecumseh Fitch. **The evolution of language**. Cambridge University Press, 2010.

[3] Charles F Hockett and Charles D Hockett. The origin of speech. **Scientific American**, 203(3):88–97, 1960.

[4] Tadahiro Taniguchi. Collective predictive coding hypothesis: Symbol emergence as decentralized bayesian inference. **PsyArXiv preprint psyarxiv.com/d2ty6**, 2023.

[5] Tadahiro Taniguchi, Yuto Yoshida, Yuta Matsui, Nguyen Le Hoang, Akira Taniguchi, and Yoshinobu Hagiwara. Emergent communication through metropolis-hastings naming game with deep generative models. **Advanced Robotics**, 37(19):1266–1282, 2023.

[6] Tadahiro Taniguchi, Shogo Nagasaka, and Ryo Nakashima. Nonparametric bayesian double articulation analyzer for direct language acquisition from continuous speech signals. **IEEE Transactions on Cognitive and Developmental Systems**, 8(3):171–185, 2016.

[7] Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, Ichiro Kobayashi, Hideki Asoh, and Masahide Kaneko. Segmenting continuous motions with hidden semi-markov models and gaussian processes. **Frontiers in neurorobotics**, 11:67, 2017.

[8] Matthew Beal, Zoubin Ghahramani, and Carl Rasmussen. The infinite hidden markov model. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, **Advances in Neural Information Processing Systems**, volume 14. MIT Press, 2001.

[9] Emily B. Fox, Michael C. Hughes, Erik B. Sudderth, and Michael I. Jordan. Joint modeling of multiple time series via the beta process with application to motion capture segmentation. **The Annals of Applied Statistics**, 8(3), sep 2014.

[10] Yasuko Matsubara, Yasushi Sakurai, and Christos Faloutsos. Autoplait: automatic mining of co-evolving time sequences. In **Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data**, SIGMOD '14, pages 193–204, New York, NY, USA, 2014. Association for Computing Machinery.

[11] Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky. The sticky hdp-hmm: Bayesian nonparametric hidden markov models with persistent states. 01 2007.

[12] Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi. Grounding of word meanings in multimodal concepts using LDA. In **IEEE/RSJ International Conference on Intelligent Robots and Systems**, pages 3943–3948, 2009.

[13] Tomoaki Nakamura, Yoshiki Ando, Takayuki Nagai, and Masahide Kaneko. Concept formation by robots using an infinite mixture of models. In **2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)**, pages 4593–4599. IEEE, 2015.

[14] Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. EGG: a toolkit for research on emergence of lanGuage in games. In Sebastian Padó and Ruihong Huang, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations**, pages 55–60, Hong Kong, China, November 2019. Association for Computational Linguistics.

図 5　実験に用いた 11 カテゴリの物体 [13]

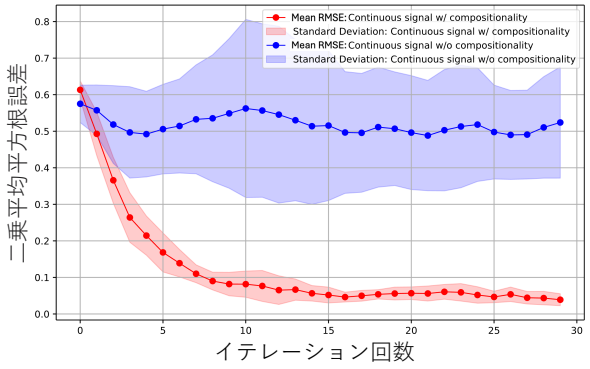

図 6　エージェント間のサインの二乗平均平方根誤差

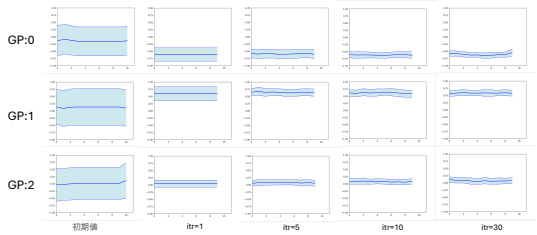

図 7　学習によるガウス過程のパラメータの推移

A　参考情報

A.1　実験設定の詳細

文献 [13] と同様，聴覚情報は 50 次元，触覚情報は 15 次元の特徴量を利用した．また，視覚情報は，同一物体の正面画像と背面画像からそれぞれ Vision Transformer で抽出した 512 次元の特徴量を各エージェントの観測情報として用いた．

MLDA における各モダリティの重みを $w_a = 5000, w_t = 5000, w_v = 500$ とし，MLDA のギブスサンプリングの反復回数を 100 回，MHNG の反復回数を 30 回とした．

A.1.1　物体を表現する連続信号の評価

特定の物体を表現する共通の連続信号が学習されているかを評価するため，67 個の物体を表現するそれぞれのエージェントの連続信号間の平均二乗誤差（MSE）を求めた．2 体のエージェントが生成した各物体を表現する連続信号が似ているほど，MSE は小さくなる．学習回数に対する MSE の変化が図 6 である．赤色の線が合成性を仮定したサインを用いた提案手法の MSE である．学習が進むに連れて MSE が低くなっており，連続信号の一致度が高くなっていっていることが分かる．一方で合成性を仮定しない手法（青線）では，初期値から MSE がほとんど変化していなく，分散も大きい．また，学習終了後のそれぞれの MSE を比較しても，約 0.55 の差があり，合成性を仮定することで一致度の高い連続信号が学習されたことが分かる．

学習の過程の連続信号の変化を確認するため，GP-HSMM の分節化の過程で各音素クラスに分類された音素波形を図 7 に示した．ランダムな初期値から始め，学習を重ねるごとに 3 種類の連続波形に収束したことが分かる．今回のガウス過程のパラメータでは，0 を平均として，おおよそ −1 ~ 1 の値の範囲の連続信号を生成しやすい設定となっていた．それぞれのエージェントが連続信号の生成と受理・棄却を繰り返すことで，−1 ~ 1 の範囲内の下限・中央・上限付近の識別しやすい 3 つの音素が学習された．