波形包絡を用いた音節核の自動抽出とそれを用いた構造的表象による単語獲得プロセスのモデル化*

☆尾崎洋輔, 峯松信明, 広瀬啓吉 (東大), Donna Erickson (昭和音楽大)

1 はじめに

クラウドやビッグデータに代表される計算機技術の発展に支えられ，音声認識技術は近年大きく進展し，様々なアプリケーションが実用化されている．その一方で認知ロボティクスの分野では，人間の言語獲得・発達を模擬するシステムを目指す研究も行なわれている [1, 2]．これらの研究では，発達心理学の知見がシステム構築において重要な役割を担っている．本研究でも幾つかの知見を参考に，幼児の言語獲得，特に単語獲得に焦点を当て，そのモデル化を検討する．

言語獲得を模擬するシステム構築を音声科学の観点から考えた場合，1) 連続する音ストリームから如何にして単語を分節するのか [3]，2) 話者・環境により大きく変化する音声を如何に正規化するのか [4]，が課題となることが多く，本研究でも後者に焦点を当てる．発達心理学の知見に基づいてこれらを検討する場合，理論・立場の違いは，異なるシステム開発戦略へと繋がる．例えば音声を音韻列として表象・操作する能力は多くの成人が有するが，幼児にこれを仮定する／しないは，その後の技術的検討に影響する [3]．

本研究では，音韻意識の発達を調査した研究例 [5] を鑑み，音韻列で表象・操作する能力を明示的に仮定しない単語獲得プロセスの模擬を試みる．この場合単語を単位として音声パターンを学習する機械を構築することになるが，話者によって異なる音声パターンを如何に一般化させるのかが問題となる．単語 HMM に代表される統計的な音声モデリング技術では，多数話者の同一単語音声を集めて分布の中に話者性を隠す技術（話者性＝隠れ変数）を構築してきた．しかし幼児の場合，聴取音声の多くは両親の声であるため，本研究の目的との整合性は低い．

話者性を（位相成分や調波成分と同様に）音声から分離し，言語的側面だけを表象することを目的として，音声の構造的表象が提唱されている [6]．本研究では上記の問題を構造表象を用いて検討する．乳児は言語のリズムに敏感に反応する様子が報告されている [7]．言語リズムは「聞こえ度」の変化パターンとして考えることができる．本研究では，音響的に定義した聞こえ度パターンから音節核を抽出し，それを用いた構造的単語音声認識の精度向上を検討する．

2 波形包絡を用いた音節核の自動推定

2.1 音節と聞こえ度

音節は一般的に母音が複数の子音を引き連れる形式で定義される（母音を V，子音を C とすると C*VC* となる）[8]．この定義の他に，音節は「聞こえ度」を用いて説明することも出来る．聞こえ度は「音声を同じ大きさ，高さ，長さで発した場合，遠くに届くものほど聞こえが大きい」と定義され，その大小関係は階層的に表される（Fig. 1 参照）[8]．これより音節は Fig. 2 のように聞こえ度の高い母音 (音節核) を中心として，聞こえ度の山谷を形作ることになる．

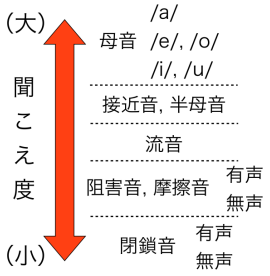

Fig. 1 音の聞こえ度

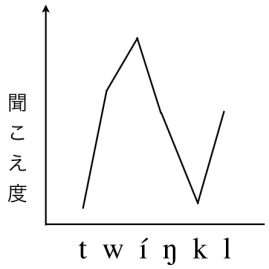

Fig. 2 聞こえ度の変化

言語リズムを考える場合，発声を一端音韻列として表現し，母音系列長や子音系列長の分散値や，発声に母音が占める割合を用いて表現したり [9]，母音系列長，子音系列長を用いて等時性を表現したり [10]，Fig. 2 に示す聞こえ度の山谷の繰り返しを言語リズムとすることもある [8]．本研究では（音韻を明示的に必要としない）この考え方に準じ，何らかの物理量を聞こえ度に対応させることで，発声のリズム構造を抽出する．具体的には，音節核の自動抽出を試みる．

2.2 先行研究

Rudi らは波形包絡情報を用いて音節情報（核・境界の位置）の自動推定を行ういくつかの手法を紹介している [11]．これらの研究は波形包絡から音節情報を直接推定しているので，波形包絡を聞こえ度に対応付けている手法と言える．

波形のエネルギーを用いて聞こえ度を推定する試みとして河合らの研究が挙げられる [12]．この研究では音声波形のフィルタバンク出力の二乗平均平方根 (Root Mean Square; RMS) から聞こえ度を推定している．音声波形の RMS 値はパワーの平均の平方根になるので，この研究ではパワーを聞こえ度に対応付けていることが分かる．

Galves らの研究ではスペクトル形状の差異を用いて聞こえ度を推定している [13]．この論文内ではスペクトル形状の差異を定式化し，スペクトル形状の変化の大きさで聞こえ度を定義している．即ち，スペクトルが時間的により安定している区間ほど聞こえ度が高いと解釈している．

---

*Extraction of syllable nuclei using waveform envelopes and structure-based modeling of infants' process of word acquisition using the syllable nuclei by Y. Ozaki, N. Minematsu, K. Hirose (The University of Tokyo), and D. Erickson (Showa University of Music)

Table 1　実験条件：音節核抽出パラメータ

BPF の帯域	500 Hz–1500 Hz
スムージング周波数	50 Hz
極大抽出区間	前後 50 ms
極大値を抽出する際の閾値	最大振幅の 1/10

Table 2　抽出精度の主観による評価

Recall	Precision	F 値
0.857	0.923	0.889

2.3　提案手法

ここではRudiらの論文[11]で紹介されているMermelsteinの手法[14]を応用した，波形包絡を用いた音節核の自動抽出の枠組みを提案する．まず，音声波形をBPF（バンドパスフィルタ）に通して周波数帯域を制限した波形を得る．得られた波形を全波整流したのちに，10 Hzから50 Hz程度のスムージング周波数でLPF（ローパスフィルタ）に通して波形の包絡を得る．ここまでの操作で得られた包絡から極大値をピックアップし，それを音節核とする．Mermelsteinの提案した方法では全ての極小値を音節境界候補としていたが，それに倣って極大値をそのまま音節核としてしまうと，不適当な音節核が大量に抽出されることになる．特に無音区間ではローパスフィルタで取り除き切れない高周波成分によって大量の挿入誤りが発生する．また，この高周波成分によって一つの山の中に複数の極大値が時間的に接近して現れることもあり，音節の定義からこれも挿入誤りとなる．/k/や/t/のような破裂音などの開始部は大きな振幅を持つため，スムージング後でも極大値として残りやすいという問題も存在した．以上の問題を解決するために，以下の方針に従って極大値を絞り込む．

- 包絡の値に閾値を設定し，閾値を越える極大値のみを音節核の候補とする．
- 前後の数十ミリ秒程度の範囲（以後，極大抽出区間）で最大であるものを選ぶ．
- 無声区間を音節核の候補から外す．

BPFの帯域やLPFのスムージング周波数，極大抽出区間長などの音節核抽出パラメータを固定して抽出すると，同じ単語でも個々の発声毎に異なる数の音節核が抽出されることがある．これとは逆に音節核抽出パラメータを適切に操作することで，所望の数の音節核位置情報を得る事も可能である．スムージング周波数を変化させると，包絡に表れるピーク数を増減できる．極大抽出区間長を伸縮させると，包絡上のピークから抽出する音節核の数を操作できる．本節の評価実験ではこの音節核抽出パラメータを固定して行うが，後述する単語認識実験では，所望する数の音節核を抽出できる特徴を利用する．

2.4　実験

音節核抽出の実験データには英文読み上げコーパスERJ[15]に含まれる米語母語話者（男性8人，女性12人）による読み上げサンプルから合計210文を無作為に抽出したものを用いた．実験条件はTable 1に示す．ここでのBPFの帯域やスムージング周波数などの音節核抽出パラメータは，予備実験に基づいて決定した．また，抽出結果の主観評価は第4著者に依頼した．ここでは第1節で示した音節の定義に捉われず，実際に音声を聴いて感覚される音節核の位置を基準に評価を行わせた．評価結果をTable 2に示す．Recall，Precision共に比較的良好な精度で抽出できているのが分かる．次節では得られた音節核情報を単語認識実験に応用する．

3　音節核情報を用いた構造的表象による孤立単語認識

前節より，音声の言語リズム情報の一種である音節核の情報が得られた．本節では音節・単語を単位とした入力音声の構造的モデリング手法[6]を提案する．

3.1　音声の構造的表象

音声の構造的表象[6]は話者性を音声から分離し，言語的側面だけを表象することを目的とした枠組みである．本節では音声の構造的表象の特徴と，これを用いた単語音声認識手法について説明する．

音声の音響特徴量に混入する非言語特徴量による歪みは，ケプストラム領域に対するアフィン変換で近似される[16]．これらの歪みの混入は不可避であるので，非言語情報によらない認識を行うにはアフィン変換に対して不変な特徴量が必要となる．線形・非線形を問わずあらゆる可逆な変換・写像に対して不変な特徴量として，分布間距離であるバタチャリヤ距離が挙げられる．分布 p_1 と p_2 に対して下記で定義される．

$$BD(p_1,p_2) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \qquad (1)$$

ここで分布 p_1，p_2 をガウス分布と仮定とすると，上の式は以下のように簡便な形に展開できる．

$$BD(p_1,p_2) = \frac{1}{8}\mu_{12}^T V_{12}^{-1}\mu_{12} + \frac{1}{2}\ln\frac{|V_{12}|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}} \qquad (2)$$

ただし，$p_1 = \mathcal{N}_1(\mu_1,\Sigma_1)$，$p_2 = \mathcal{N}_2(\mu_2,\Sigma_2)$ であり，$\mu_{12} = \mu_1 - \mu_2$，$V_{12} = (\Sigma_1 + \Sigma_2)/2$ とする．ケプストラムベクトル系列を K 個のガウス分布の時系列で表現されるとすると，バタチャリヤ距離によって張られる $K \times K$ の距離行列が得られる．バタチャリヤ距離はアフィン変換に対して不変なので，この距離行列は非言語特徴量に対して凡そ不変な音響的特徴量となり，これを音声の構造的表象と呼ぶ．

上記で得られた $K \times K$ の距離行列は対角成分が0となる対称行列である．対角成分を除いた上三角成分の値を並べた $K(K-1)/2$ の次元を持つベクトルを構造ベクトルと呼び，これを入力発声の特徴ベクトルとして使用する．単語認識実験では，Fig. 3のように構造統計モデルを多次元ガウス分布と仮定して，学習データからこれを推定する．認識の際には入力音声の構造ベクトルがこの構造統計モデルから出力

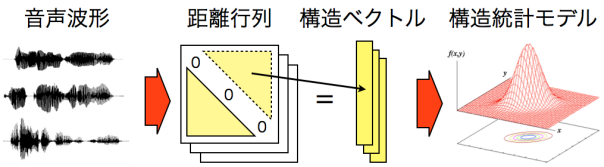

Fig. 3　一単語の構造統計モデルの作成

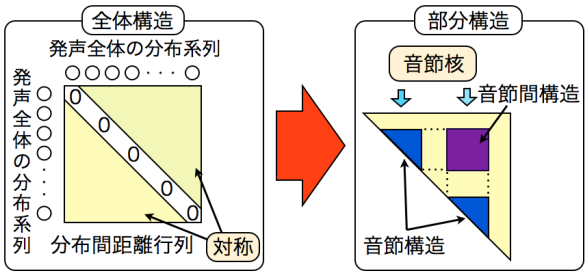

Fig. 4　部分構造の定義

される確率を尤度スコアとして認識を行う．

3.2　提案手法

本節では，音節核の位置情報を手掛かりとして，音節単位で構造的表象によるモデリングを行う手法を提案する．発声全体を分布系列化した際に，各分布の時間情報より音節核が所属する分布を決定できる．音節核が所属する分布を中心に長さ $K_s(3 \leq K_s \leq 5)$ の分布系列を取り出し，その分布系列内で計算される距離行列を「音節構造」と定義する．更に異なる音節核から得られる分布系列の間で計算される距離行列を「音節間構造」と定義する．前節で説明した単語発声全体の構造的表象を「全体構造」と呼ぶこととし，音節構造と音節間構造をまとめて全体構造に対する「部分構造」と定義する．全体構造と部分構造の様子を Fig. 4 に示す．部分構造では，発声毎に抽出された音節核の数によって得られる距離行列の数が異なる．発声中の音節数を N とすると N 個の音節構造と，${}_NC_2$ 個の音節間構造が得られる．

前節で述べた全体構造の統計モデル同様，部分構造の統計モデルを作成する．部分構造の枠組みでは得られる音節核の数によって得られる音節構造・音節間構造の数が変化する．よって同一単語であっても，発声毎に抽出される音節核数が変化するのは望ましくない．そこで音節核抽出パラメータを固定せず，所望の数の音節核が得られるようパラメータを可変化して抽出する．学習に用いる音声の単語情報は既知と考え，単語毎に抽出すべき音節核数を決定する．

学習音声と異なり入力音声の単語情報は未知なので，予め抽出する音節核の数を決定することが出来ない．本実験で用いたデータには 2 音節単語から 5 音節単語までしか存在しないので，音節核数を 2〜5 と仮定して各々部分構造を計算し，これらを使い分けながら各単語モデルとの照合を行う．

Table 3　音響分析条件

サンプリング	16 bit / 16 kHz
窓	25 ms length / 10 ms shift
特徴量	MFCC 12 次元 + ΔMFCC
分布推定方法	MAP 推定 [17]
出力分布	対角共分散ガウス分布
状態数	20

Table 4　音節核抽出条件

BPF の帯域	500 Hz-2500 Hz
スムージング周波数	10 Hz から
極大抽出区間	スムージング周波数毎に全探索
極大値を抽出する際の閾値	最大振幅の 1/10
無声区間	カットしない

3.3　尤度スコアの統合

全体構造による対数尤度スコアに対して，部分構造に基づく平均対数尤度スコアを重み付けして計算される統合スコアを用いて孤立単語認識実験を行う．発声全体の構造的表象を用いて計算された対数尤度スコアを S，n 番目の音節構造の対数尤度スコアを T_n，m 番目の音節間構造の対数尤度スコアを U_m とすると，統合されたスコア V は下記の式で表される．

$$V = S + \omega \left(\frac{1}{N} \sum_{n=1}^{N} T_n + \frac{1}{M} \sum_{m=1}^{M} U_m \right) \qquad (3)$$

ただし M は $M = N(N-1)/2$ で表される音節間構造の総数である．この V を最大化するモデルが認識結果となる．実験では発声全体の構造で照合を行った場合のスコアに対する，部分構造のスコアの重み ω を変化させて実験を行う．

3.4　実験

単語認識実験のデータベースには東北大-松下 単語音声データベースの Set1: 音韻バランス 212 語（男性 30 名，女性 30 名の 60 話者）を使用した．各発声における分布の推定，音節核の抽出はそれぞれ Table 3, Table 4 の条件で行う．また構造的表象の強すぎる不変性を抑えるためにケプストラムの部分空間を用いたマルチストリーム化 [6] を行っている．各ストリームの次元数であるブロック長 L は 12（次元分割を行わない），2，1 の 3 種類を用いた．構造の計算は MFCC のみを用いて行い，部分構造を計算する際の分布系列長 K_s は 3，5 の場合の 2 種類で実験を行った．

認識実験の結果を Fig. 5 に示す．マルチストリーム化のブロックサイズ L が小さいほど全体的な認識率が向上していることが分かる．ここで ω=0 の場合は，全体構造のみを用いて認識を行うことになるので，従来の構造的表象のみを用いた場合の認識精度となる．部分構造の重み ω が増加するに従って認識精度は上昇するが，ある程度部分構造の比重が大きくなってくると認識精度は減少し始め，最終的には ω=0 の場合を下回る．また，部分構造に用いる分布系列数が増えると精度向上の割合が増大しているのが分かる．今回の実験の範囲内で最も高い認識率は部分構造の分布系列数が 5 で，L=1，ω=2 の時の 84.0%で，ω=0

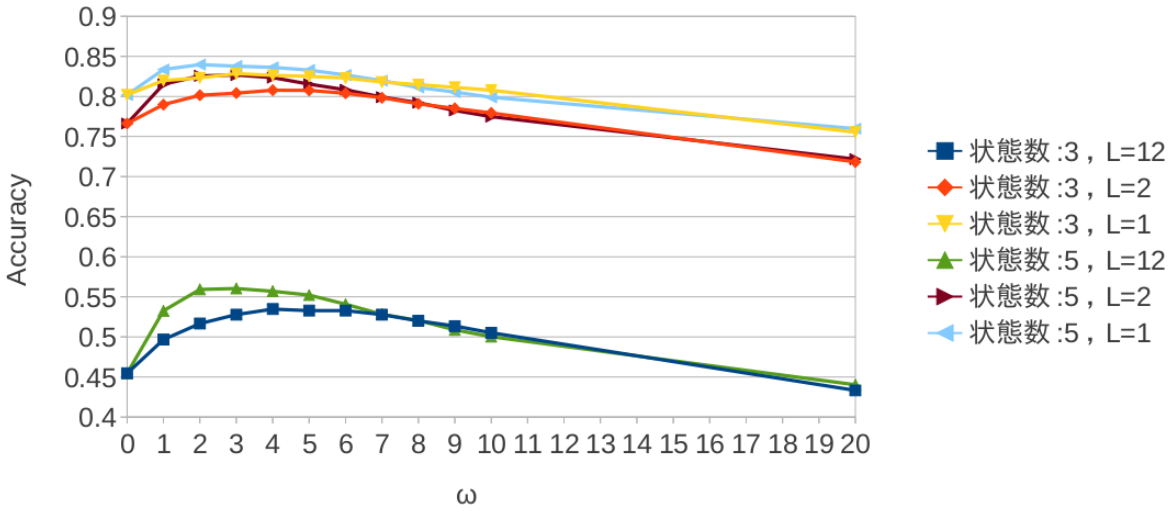

Fig. 5　孤立単語認識実験の結果

の場合と比べて誤り削減率は 19.1%となっている．

適切な ω を設定する事によって大きな精度向上が得られた事からも，部分構造を用いて音節単位でのモデリングを行うことが有効であることが分かる．

4　まとめと今後の課題

明示的に音韻の存在を仮定せず，話者性を排除して単語音声全体を表象する音声の構造的表象に基づく単語音声認識系を用いて，幼児の単語獲得プロセスの模擬を検討した．言語リズムに敏感な幼児の特性を考慮し，言語リズムと関連が深い音響量を用いて構造的な単語音声認識の高精度化を試みた．

前半の実験では，波形包絡が「聞こえ度」に凡そ対応していると考え，波形包絡からの音節核自動抽出を行った．その結果，比較的良好な音節核の自動抽出を実装することができた．後半では，音節核情報を手掛かりに，音節単位での構造的モデル（部分構造）を提案し，これらを用いた構造的孤立単語認識の高精度化を試みた．全体構造と部分構造を適切な重みで統合することにより，認識率の改善を測ることができ，音節核位置に基づく部分構造の有効性を確認できた．

今後の課題としてはまず第一に音節核の自動抽出の高精度が挙げられる．本稿では最も単純な Mermelstein の手法を用いたが，フィルタバンクを用いたり今回紹介した複数の手法を組み合わせて実装を行う事により更なる高精度化が期待出来る．また，母親が幼児に話しかけるような言葉 (motherese) を実験に用いた場合に精度がどう変化するのかについても検討する予定である．この場合は言語リズムがより捉え易くなっており，音節核推定精度と部分構造を用いた単語認識精度の両方が向上することが期待される．

また，本稿では音声の連続ストリームに対する分節問題は，一切取り扱わなかった．音声ストリームから，頻出する音声パターンを抽出する課題は教師無しパターン発見（Unsupervised pattern discovery）として幾つかの検討例 [18] がある．構造的な特徴を用いたパターン検出についても検討する予定である．

参考文献

[1] ACORN Project: http://lands.let.ru.nl/acorns/index.htm

[2] Developmental Speech Recognition Workshop: http://www.cit-ec.de/workshop_developmental_speech_recognition

[3] O. Räsänen, “A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events,” Cognition, 120, 149–176, 2011.

[4] G. Ananthakrishnan *et al.*, “Using imitation to learn infant-adult acoustic mappings,” Proc. INTERSPEECH, 765–768, 2011.

[5] 原恵子, “子どもの音韻障害と音韻意識”, コミュニケーション障害学, 20, 2, 98–102, 2003.

[6] 峯松信明他, “音声に含まれる言語的情報を非言語的情報から音響的に分離して抽出する手法の提案　～人間らしい音声情報処理の実現に向けた一検討～”, 電子情報通信学会論文誌, J94-D, 1, 12–26, 2011

[7] R. Mazuka, “The Rhythm-based Prosodic Bootstrapping Hypothesis of Early Language Acquisition:Does It Work for Learning for All Languages?” GENGO KENKYU, Vol. 132, pp. 1–15, 2007.

[8] 窪薗晴夫, “音声学・音韻論”, くろしお出版, 1998.

[9] F. Ramus, “Autoamtic correlates of linguistic rhythm: Perspective,” Proc. Speech Prosody, 2002.

[10] E. Grabe *et al.*, “The acquisition of rhythmic patterns in English and French,” Proc. ICPhS, 1201–1204, 1999.

[11] R. Villing *et al.*, “Performance Limits for Envelope based Automatic Syllable Segmentation,” IEE Irish Signals and Systems Conference, 2006.

[12] G. Kawai and J. V. Santen, “Automatic detection of syllabic nuclei using acoustic measures,” IEEE Workshop on Speech Synthesis, 2002.

[13] A. Galves *et al.*, “Sonority as a basis for rhythmic class discrimination,” Speech Prosody 2002, Aix-en-Provence, 2002.

[14] P. Mermelstein, “Automatic segmentation of speech into syllabic units,” Journal of the Acoustical Society of America, Vol. 58, pp. 880–883, 1975.

[15] 峯松信明他, “英語 CALL 構築を目的とした日本人及び米国人による読み上げ英語音声データベースの構築”, 日本教育工学会論文誌, 27, 3, 259–272, 2004.

[16] M. Pitz *et al.*, “Vocal tract normalization equals linear transformation in cepstral space,” IEEE Trans. Speech Audio Process., 13, 5, 930–944, 2005.

[17] 村上隆夫他, “音声の構造的表象に基づく日本語孤立母音系列を対象とした音声認識”, 電子情報通信学会論文誌, J91-A, 2, 181–191, 2008.

[18] A. S. Park *et al.*, “Unsupervised pattern discovery in speech,” IEEE Trans. audio, speech, and language processing, 16, 1, 186–197, 2008.