修士論文

乳児の言語獲得プロセスのモデル化に基づく音声処理技術の検討
(A study of speech processing technology based on modeling the processes of infants' language acquisition)

2014 年 2 月 6 日

指導教員　峯松 信明 教授

電子情報学専攻

48-126411　尾崎 洋輔

内容梗概

乳児の言語獲得プロセスを計算機上で再現しようという試みが行われるようになって久しく，現在では認知ロボティクスの分野だけでなく発話内容（言語）の不明な音声／動画に対するキーワード抽出，未知言語に対するモデル構築など，その応用も多岐に渡っている．従来，そのような研究の多くは脳神経科学や発達心理学などの研究によって得られた知見を参考に，入力された音声ストリームからボトムアップに音韻や語彙を獲得するものが主流であった．しかし，実際には言語知識などから経験的にパラメータチューニングを行う必要があったり，複数話者による音声データに対しては途端に頑健性を失う（教師無しパターン抽出における複数話者問題）など問題もいくつか存在した．

そこで本研究では発達心理学に基づき，言語獲得プロセスを模擬するシステムの特徴を以下のように定義して，実際にこれらを満たす音声処理について実験・検討を行った．

- 連続音声中から言語や音韻などの事前知識を用いずにボトムアップに語彙を獲得する．
- 言語リズム[1]に即した音声処理を行っている．
- 単語全体の語形である語ゲシュタルトを捉えており，非言語情報の変化に頑健である．

まず，音声波形からボトムアップに言語リズムの抽出する方法について評価実験を行い，その結果を用いて言語リズムに即した語ゲシュタルトのモデル化を行った．音声波形からの言語リズムの抽出に関しては特に主観評価でとても高い精度を示し，提案手法が人間の知覚により即した推定が行えている事を示した．また後者の実験は語ゲシュタルトの物理的定義として峯松らの構造的表象を用いており，音節情報を用いた構造的表象の拡張とも言える．音声の構造的表象は非言語情報に対して普遍的構造を持つ一方で，音声事象間の厳密な対応が必要となる時間アライメントの問題が存在し，言語リズムの情報はこの問題に対する一つの解決策であり，実際にこのモデルを用いて孤立単語認識タスクで音声認識を行って認識精度の向上を示した．

次に語ゲシュタルトを考慮した連続音声からの教師無し語彙獲得について実験を行い，その効果を検討した．上述した通り従来の教師無しパターン発見タスクにおいては複数話者音声への頑健性が一つの問題となっている．これはラベル付き音声が学習データとして提供される一般的な音声処理とは異なり，非言語情報が多分に含まれる空間上で異なる話者の同一内容発話を比較する必要があるからである．語ゲシュタルトの物理的定義である音声の構造的表象を音声照合の際に用いる事で，この問題については一定の成果が得られたが，他の音声正規化手法に対する明確な有意差は得られていない．

[1] 本研究では音節の出現間隔でこの言語リズムを定義している

目次

図目次

表目次

第1章

序論

1.1　背景

人間の重要なコミュニケーション手段である音声が研究・分析の対象となって久しく，今日では音声処理技術の発展や計算機能力の飛躍的な向上により，手元の端末でも簡単に音声認識を行えるまでに至った．現在主流になっている大語彙音声認識システムでは，音声はその構成単位のひとつである音素の形にまで分解され，個々の音素ごとにモデル化を行うことにより認識が行われる [1]．この手法では，膨大な音声サンプルを使用した統計学的なアプローチによって，話者性などの特徴の変化を隠れ変数として扱って不特定多数の話者の音声を認識できるように音声をモデル化する．しかし，このような認識システムは話者性の大きな変化までは対応し切れない場合もある．この場合，音声特徴量を正規化したり，音響モデルパラメータを適応させるなどして対処している．しかし，現在の音声技術が本当に人間の知覚に即したものであるとは言えない．

こうした認識精度の向上が主な目的となる技術とは異なり，音声認識システムと人間の言語的発達を関連付けて，より人間らしい音声認識技術を構築しようとする研究も行われている [2, 3, 4, 5]．これらの研究では，音声認識システムがモデルを学習して発声を認識する過程を，幼児が言語を獲得する過程に当てはめている．その応用先も認知ロボティクスの分野だけでなく，未知語の検出や未知言語に対するモデル構築など多岐に渡っており，工学的にも深く研究されている分野とも言える．本研究でも発達心理学の見地によるいくつかの報告に基づき，言語獲得プロセスを模擬するシステムの特徴を以下のように定義した．

- 連続音声中から言語や音韻などの事前知識を用いずにボトムアップに語彙を獲得する．
- 言語リズムに即した音声処理を行っている．
- 単語全体の語形である語ゲシュタルトを捉えており，非言語情報の変化に頑健である．

本研究の最終的な目標は上記の特徴を兼ね備えた言語獲得フレームワークを構築する事である．

従来では，それぞれの観点において

- 連続音声ストリームからの教師無しパターン抽出,
- 音声波形からの言語リズム（音節情報）抽出,
- 非言語情報に対して普遍的構造である音声の構造的表象,

のような先行研究（領域）がほぼ独立して存在していた．しかし，実際には上記の特徴は独立ではなく，それぞれが相互に影響を与え合っていると思われる．例えば事前知識が一切仮定できない乳児にとって，言語リズムは音声ストリームを分析する為の重要なマーカーであると考えられる．また教師無しで語彙を獲得する際には，発話者が異なっていたとしても発話内容が同じであれば，それらを同一のものだと認識するメカニズムが必要となる．そのために上記のような手法をいくつか組み合わせたフレームワークを提案し，それらの有用性を実験で検証する．

1.2　本論文の構成

本論文は全 6 章から構成される．まず第 1 章では，本論文の背景と目的について述べる．第 2 章では，従来の工学的観点における音声処理の枠組みと，それを実現する要素技術について説明する．第 3 章では，乳児の言語獲得過程について発達心理学の観点からの研究事例を紹介し，どのような特徴を持つのかを考える．ここでは特に (1) 乳児の言語的発達のプロセス，(2) 音声コミュニケーションのベースとなる言語のリズム（音節リズム），(3) 乳児（人間）が音声認識時に捉える音声の特徴，に焦点を当てて解説する．第 4 章では，上記の特徴の中でも言語のリズムに焦点

を当て，教師無しでの言語リズムの抽出とそれを用いた音響モデリングについて実験・検討する．まずは音声波形から言語リズム（音節情報）の自動抽出する実験を行い，そうして得られた音節情報を用いて，言語リズムに即した音響モデルを提案する．第 5 章では，語ゲシュタルトの概念を考慮し，複数の話者が存在する入力音声信号に対しても頑健に動作する教師無し語彙獲得について実験を行う．ここでは従来の教師無しパターン発見アルゴリズムに，語ゲシュタルトの物理的定義である構造的表象を適用する事で提案手法となる．第 6 章で本論文をまとめ，今後の展望について述べる．

第2章

従来の音声認識の枠組み

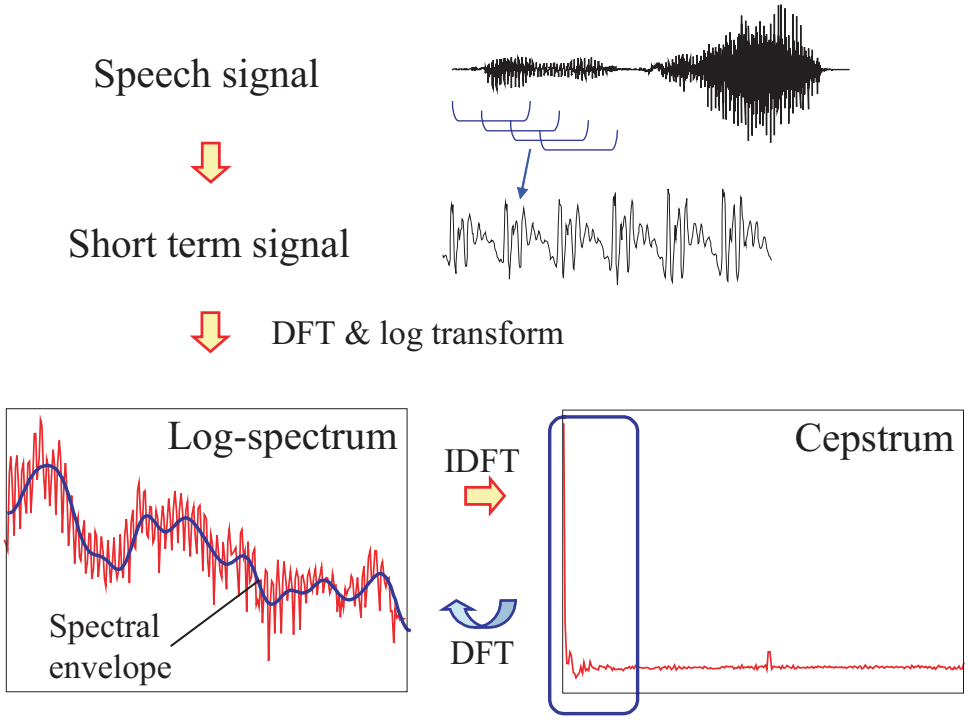

図 2.1: ケプストラムの抽出

2.1　はじめに

本章ではまず従来の音声認識のアプローチについて紹介し，それらの技術が人間の知覚や言語獲得と，どのような関係にあるのかを簡単に説明する．また本研究では乳児の言語獲得をテーマとしているので，音声処理の中でも特に，(1) 現在の音声システムは音のどのような特徴を捉えているのか（音声分析・音響特徴量），(2) 時間方向に柔軟に伸縮し話者によってのバラつきの大きい音声をどのようにモデル化するのか（音声モデリング），という部分に焦点を当てて解説する．

2.2　音響特徴量

2.2.1　ケプストラム

HMM を用いた音声認識においてよく用いられる特徴量としてケプストラム (Cepstrum) が挙げられる．音声からのケプストラムの抽出の流れを図 2.1 に示す．音声波形とそのスペクトルは時間的に変化するものであるが，数十ミリ秒単位のフレームを切り取って見れば，フレーム内のスペクトルは定常状態と見なすことができるので，音声波形を数十ミリ秒のフレームに切り分けてフレームごとに対して処理を行う．まず，切り取ったフレームに対して短時間フーリエ変換を施してスペクトルを抽出し，得られたスペクトルを対数パワースペクトルに変換し，逆離散フーリエ変換を施す．こうして得られたデータ列の低次項にはスペクトルの包絡特性が，高次項にはスペクトルの高周波成分の特性が現れていることになる．このスペクトルの高周波成分は音の基本周波数[1]となっており，主に非言語情報である話者情報やパラ言語情報を表す．一方，スペクト

[1]話者の声の高低．基本周波数の変動により感情などのパラ言語情報を表すこともある．

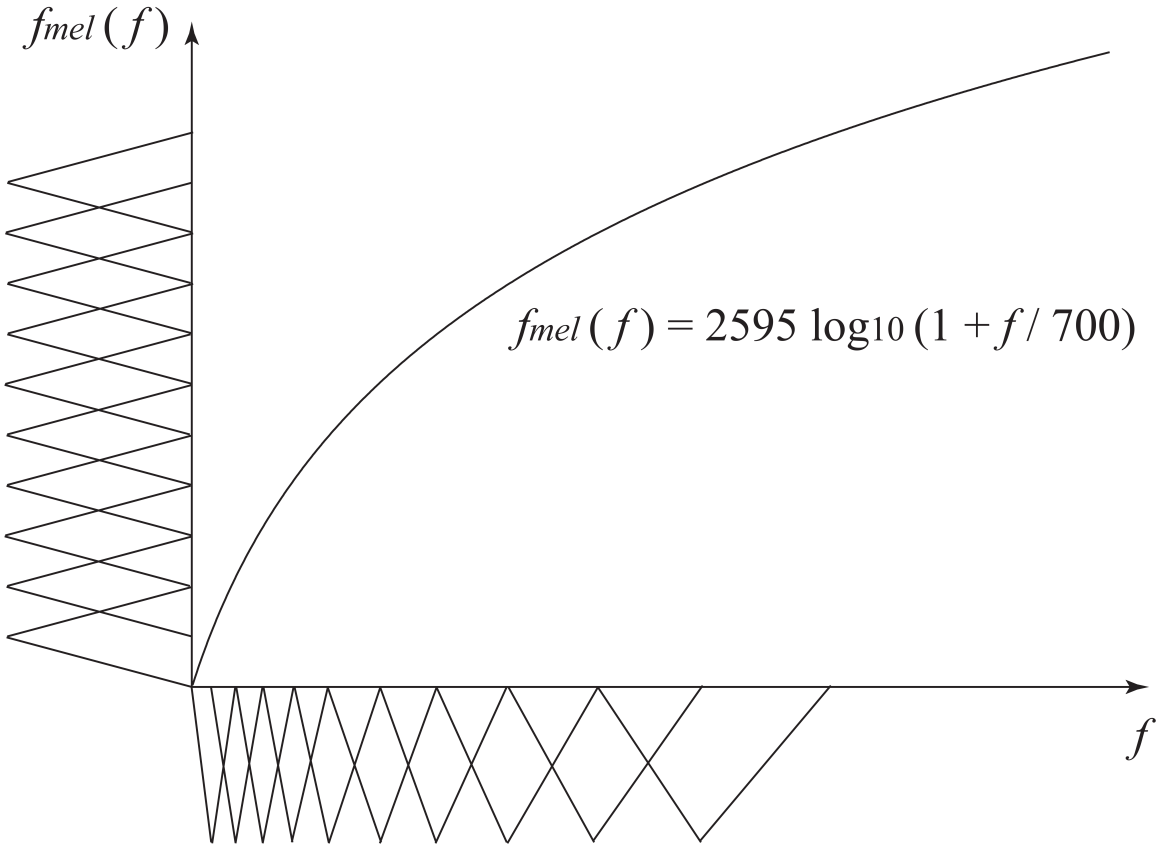

図 2.2: メル周波数とその軸上に等間隔で配置された三角窓

ルの包絡成分には音声の音韻的特徴がよく表れるとされている [6]．この低次項のみを取り出したものをケプストラムと呼び，ベクトルの形で抽出される．この操作をフレーム長の半分程度のシフト長でフレームシフトしながら繰り返すことにより，フレームの数だけのケプストラムベクトルの系列が得られる．

ケプストラムはその抽出過程からも分かるとおり，音声の音韻的特徴を小数のパラメータで効率よく評価することのできる特徴量である．また，ケプストラムは次元間の相関がほとんど無いことが知られており，ケプストラムに対する様々な計算を簡便に行うことができる．

2.2.2　メル尺度に基づくケプストラム特徴

人間の周波数分解能は低い周波数ほど細かく，高周波ほど粗くなり，その特性は音の周波数の高さに対してほぼ対数の関係で表れる事が知られている．この人間の音の高さに対する感覚はメル尺度と呼ばれており、ケプストラムにこの尺度を反映させた特徴がいくつか提案されている．ここで紹介する MFCC (Mel-Frequency Cepstrum Coefficient) もその一つである．MFCC の導出ではまず，図 2.2 のようにメル周波数 (メル尺度化された周波数) 軸上に等間隔で配置された三角窓を用意し，フィルタバンク分析を行なう．ここでメル周波数 f_{mel} は以下の式のように周波数 f にウォーピングを施す事で得られる．

$$f_{mel}(f) = 2595\log_{10}(1 + \frac{f}{700}) \tag{2.1}$$

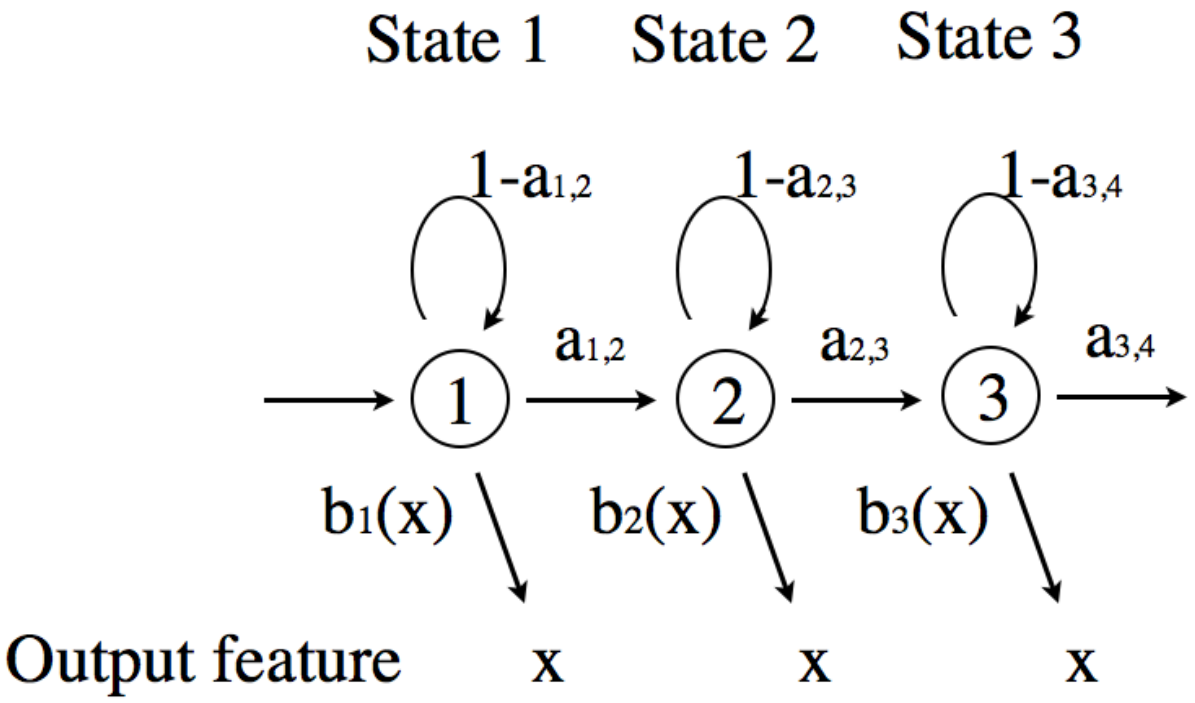

図 2.3: 隠れマルコフモデル

各窓に対応する周波数帯域のパワースペクトルを，窓の大きさによる重み付けして和を計算する事でメルスペクトルが得られる．このメルスペクトルを離散コサイン変換する事により MFCC が得られる．

2.2.3　ケプストラムの動的特徴

スペクトル（ケプストラム）の時間的な変化も特徴として用いられる事も多い．音声分析によって得られるケプストラムベクトル時系列を，時間方向に線形近似した傾きを Δ ケプストラムと呼び，ケプストラム特徴に連結して用いられる．現在は Δ ケプストラムの動的特徴である ΔΔ ケプストラムなども特徴として使用される事が多い．

2.3　音声のモデル化

2.3.1　隠れマルコフモデル

音響モデル[2]とは，ある単語 W を意図して発声した場合にどのような音響特徴 X が出力されるかの確率 $P(X|W)$ が記述されているモデルである．音声の時間方向に柔軟に伸縮するという特性から，音響モデルには図 2.3 のような隠れマルコフモデル (Hidden Markov Model; HMM) を用いるのが一般的である．隠れマルコフモデルでは，観測されるデータがどの状態から生じたのかは観測されず，隠れ変数となっている．ここで $b_1(x)$，$b_2(x)$，$b_3(x)$ は各状態から特徴が出力され

[2]音声は音の特徴による音響モデルと言語構造などの情報による言語モデルの 2 つのモデルの組み合わせによって表される．

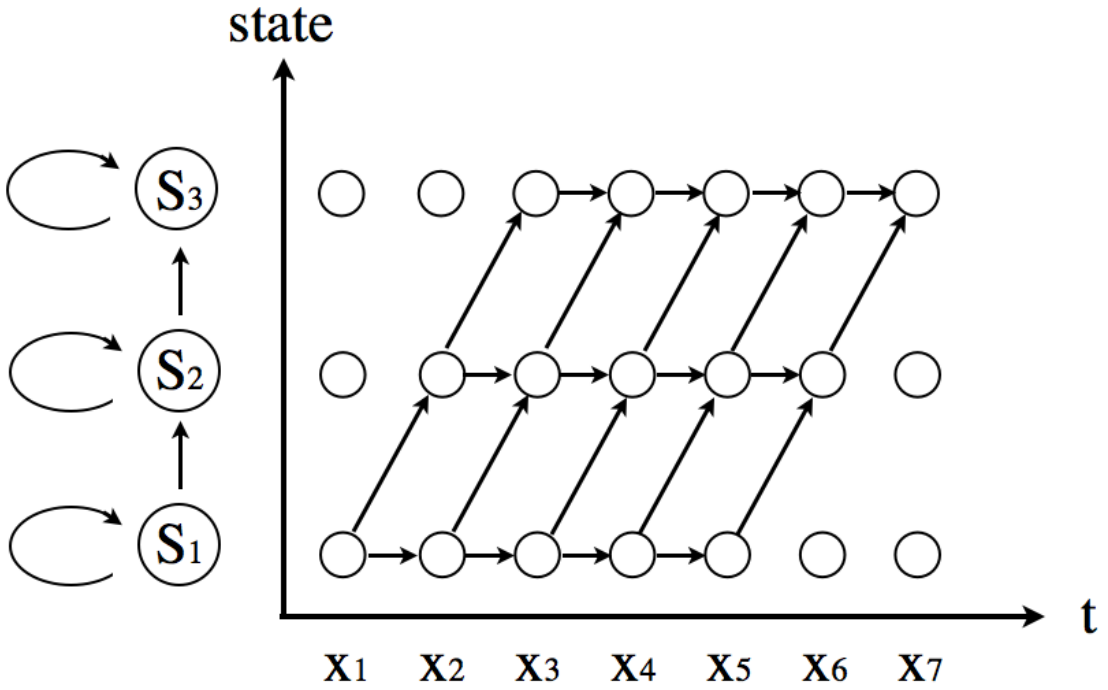

図 2.4: HMM の状態遷移の経路

る確率（出力分布）を表し，$a_{1,2}$，$a_{2,3}$，$a_{3,4}$ は次の状態への遷移確率[3]である．HMM は各状態が一定回数だけ出力分布 $b_i(x)$ に従ってフレーム特徴を出力してから，次の状態に移るという動きを繰り返すオートマトンであり，音響特徴の生成モデルとも呼ばれる．また，出力分布 $b_i(x)$ としては，複数のガウス分布の重み付け和で一つの分布を表現する混合ガウス分布 (Gaussian Mixture Model; GMM) がよく用いられる．

2.3.2　HMM を用いた音声認識

まずは HMM を用いて音声認識を行う方法について，孤立単語認識の例で説明する．HMM にはモデルから音響特徴 X が出力される確率 $P(X|W)$ が記述されているので，この場合は出現する単語の数だけ HMM を用意しておき，観測された特徴が出力される確率（尤度）が最も高い HMM が認識結果とすればよい．図 2.4 は，観測された音響特徴量 $X = (x_1, x_2, ..., x_7)$ が，ある単語 HMM から出力される場合の可能な状態遷移を表している．ある一つの経路を通り X が出力される確率は，各状態で x_i が出力される確率 $b(x_i)$ とそれぞれの経路において選択された遷移の場合の確率（$a_{i,i+1}$，$1 - a_{i,i+1}$）の積で表される．全ての可能な経路においてこの出力確率の計算を行い，その総和をとる事でその単語 HMM から観測信号 X が出力される確率 $P(X|W)$ となる．また全ての経路の場合を計算する事で計算量が増大してしまうので，実際には出力確率が最大となる[4]経路のみを計算し，その経路における出力確率で $P(X|W)$ を近似している（ビタビアルゴリズム）．連続音声認識や大語彙音声認識では全ての出現する可能性のある文章をモデル化する事

[3]自己遷移確率はそれぞれ $1 - a_{1,2}$，$1 - a_{2,3}$，$1 - a_{3,4}$ となる．

[4]出力確率が最大となる点を厳密に求めようとすれば計算量は変わらないので，実際には適切な枝刈りなどが行われている．

は不可能なので，音素や単語と言った文章のサブセットの単位で HMM を用意しておき，それを連結する事で単語モデルを構築している．ここで前後の音との調音結合を考慮しない音素 HMM は mono-phone，考慮する場合では音素 tri-phone と呼ばれる．

2.3.3　HMM の学習

HMM において推定すべきパラメータ θ は $\{a_{i,i+1}, b_i(x)\}$ であり，一般的に最尤 (Maximum Likelihood; ML) 推定によって求められる．これはある HMM の学習データとして音響特徴量の時系列 X が観測された時に，そのモデルに対する尤度を最大化するという問題に帰着できる．

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(X|\theta) \tag{2.2}$$

しかし上式を解析的に解くのは困難であるため，実際には局所最適解を求める Baum-Welch アルゴリズムが用いられる．

Baum-Welch アルゴリズムではまず，以下の式で表される前向き確率 $\alpha_i(t)$，後向き確率 $\beta_i(t)$ を各格子点毎に計算する必要がある．

$$\alpha_i(t) = P(z_{ti} = 1, x_1, , x_t|\theta) \tag{2.3}$$

$$\beta_i(t) = P(x_{t+1}, , x_T|z_{ti} = 1, \theta) \tag{2.4}$$

ここで z_{ti} は時刻 t における状態が i であれば 1，そうでなければ 0 をとる隠れ変数である．この前向き確率 $\alpha_i(t)$ と後向き確率 $\beta_i(t)$ を用いて，時刻 t における状態が i である確率 $\overline{z_{ti}}$ を

$$\overline{z_{ti}} = P(z_{ti} = 1|X, \theta) \tag{2.5}$$

$$= \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \tag{2.6}$$

のように表せる．これによりパラメータ θ と学習データからの特徴時系列 X が用意されれば，特徴系列の各々が特定の状態から生じた確率が求まり，モデルパラメータの更新が行える．例えば，出力確率 $b_i(x)$ の分布として単一ガウス分布 $(x; \mu_i, \sigma_i^2)$ を用いる場合には，パラメータ $\theta = \left\{a_{i,i+1}, \mu_i, \sigma_i^2\right\}$ から，新しいパラメータ $\hat{\theta} = \left\{a_{i,\hat{i+1}}, \hat{\mu_i}, \hat{\sigma_i^2}\right\}$ を以下のような式によって最尤推定できる．

$$a_{i,\hat{i+1}} = \frac{\sum_{t=1}^{T-1} \alpha_i(t) a_i b_i(x_{t+1}) \beta_{i+1}(t+1)}{\sum_{t=1}^{T-1} \alpha_i(t)\beta_i(t)} \tag{2.7}$$

$$\hat{\mu_i} = \frac{\sum_{t=1}^{T} \overline{z_{t,i}} x_t}{\sum_{t=1}^{T} \overline{z_{t,i}}} \tag{2.8}$$

$$\hat{\sigma_i^2} = \frac{\sum_{t=1}^{T} \overline{z_{t,i}} (x_t - \mu_i)^2}{\sum_{t=1}^{T} \overline{z_{t,i}}} \tag{2.9}$$

このようにして得られる新たなパラメータ $\hat{\theta}$ と θ の間には

$$P(X|\hat{\theta}) \leq P(X|\theta) \tag{2.10}$$

のような関係が成り立ち，パラメータの更新後は更新前と比べて学習データの尤度が上昇する事が保証されている．ただし，Baum-Welch アルゴリズムは飽くまでもモデルパラメータの更新によって局所最適解を探し出すアルゴリズムなので，ある程度適切なパラメータを持った初期モデルが必要となる．

2.4　従来の音声技術と人間の知覚との関係性

本章では従来の音声処理技術の中でも，特に音響分析と音響モデルについて説明を行った．音声のどの側面に必要な情報が表れるのかは，人間の知覚の際に音声のどのような特徴を捉えているのかという事が重要な手掛かりとなっている．音声信号は空気の振動であり観測されるのは波形であるが，実際の音声分析はスペクトルやその派生特徴から行われる．これは音声信号中の情報が短時間スペクトルの形状によく表れるからであり，人間の耳に特定の周波数に反応する器官（コルチ器）が多数配置されていて，周波数毎に分解されて脳に電気信号が伝えられるというメカニズムを持つ事が根拠となっている．このように人間の知覚に即した特徴を用いる事で効率良く音声の特徴を抽出する事ができ，メル尺度を用いた MFCC なども正にその考えに依る．

しかし，MFCC のような特徴が果たして本当に人間が捉えている特徴なのかは疑問が残る．現在の音声技術では音声の言語情報と非言語情報（話者の違いなど）を完全に切り離す事は出来ず，不特定多数の話者の音声を認識できるようにする為に，大規模なラベル付き学習データに統計学的なアプローチを用いて対処をしている．HMM の出力分布として複数のガウス分布から成る GMM を用いて様々な話者の発話を表現したり，個々の話者に応じて音響モデルパラメータを調節する話者適応などは，その代表的なものである．しかしながら，前者のアプローチで吸収できる話者性の違いには限界があり，後者についてもあらかじめ学習データを十分に用意しておく必要があるなど，人間の頑健性に遥かに及ばない．言語情報や背景などをあらかじめ仮定する事のできない乳児が，限られたリソース（例えば音声提供者が両親，兄弟しか居ない）から頑健な音声コミュニケーション能力を獲得する点，特に彼らが声帯模写では無く「彼ら自身の声で言葉を発する事が出来るようになる」という様子から，人間が音声を捉える際に事前知識に寄らず言語情報をそのものを捉えている事を示唆している．

本章で紹介した技術はいずれも工学の立場によって提案・発展してきた手法である．次章では発達心理学の知見などから，人間らしい音声処理とはどのようなものであるかを考察する．

第3章

発達心理学の知見による人間らしい言語処理

3.1　はじめに

前章では現在の音声処理で用いられている要素技術について説明し，それらと人間の知覚・言語獲得との関係を議論した．本章では逆に，発達心理学の立場から人間らしい言語獲得・音声処理とはどのようなものなのかを考える．特に乳児の言語獲得を実現する上で重要となる，(1) 乳児はどのような手順で音声を理解し言語を獲得するのか，(2) 乳児は音声をどのように知覚しているのか，(3) 乳児は音声の言語情報をどのように捉えているのか，という部分について本論文の立場を決定し，関連研究をいくつか紹介する．

3.2　教師無し学習による語彙獲得

言語獲得の最初期段階における最大の問題として，言語知識を一切持たない乳幼児が周囲から得られる連続音声ストリームから如何にして語彙を獲得するかという問題が挙げられる．乳幼児にとっては外界から得られる音声は全て未知語の集まりであり，またそれらの知覚単位がどのようなものなのかまでは明確ではない．言語能力が十分に発達した人間ならば，既存の語彙情報や前後の文脈などによって未知語入力を検知し，それを音韻列で解釈する事で新たな語彙の獲得を行う事が出来る．また一般的な音声認識システムでは，あらかじめ用意されたサンプルを基に (1) 音素や音節といったサブワード単位でのモデリングを行う，(2) サブワードの組み合わせとして語彙を大量に保持する，という方法でこのような問題に対処している．しかし，いずれの場合も語彙知識や言語処理能力のバックグラウンドが前提となっている．一方，このような先天的な言語能力を新生児に仮定しない[1]とすると，言語獲得プロセスのシミュレーションにおいて前述したようなアプローチは不可能になる．一般的な音声システムにおいて用意される語彙情報、音韻情報（音素や音節も含む）、そしてそれらの語彙や音韻に紐付けられた音声サンプルなどを、外界から入力される音声信号のみから構築しなければならない。

上記の問題は以下の３つのタスクに分解できる．

- 音韻系列の学習
- 連続音声から語彙トークン[2]へのセグメンテーション
- 語彙トークンの意味によるセグメントのカテゴライズ

ここで音韻意識が先か語彙形成が先かという問題が生じる。我々成人が語彙を獲得する際には，音声を音韻の連なりとして解釈する事からも[3]，まず音韻情報が獲得されてそれから語彙の獲得が始まるように思えるが，発達心理学の立場ではこれを無条件には了承せず，この問題に対する明確な解答は用意されていない．そのため，この問題へのアプローチについて，音声工学の立場からは２つの代表的な仮説が存在する．１つめの仮説は Kuhl らが論文内で NLM-e (Native Language Magnet theory expanded) と呼ぶ理論で、乳幼児は周囲の音声からまず始めに音韻情報を形成し、それから語彙を獲得して行くという仮説となっている [7]．２つめの仮説は Werker らが論文内で PRIMIR (a developmental framework for Processing Rich Information from Multi-dimensional Interactive Representations) と呼ぶフレームワークで，こちらでは語彙の形成が音韻意識に先立って行われている．まず連続音声からプリミティブな語彙を獲得し，それから内部の音韻連結構造を学習する．そして新たに得られた音韻情報によって更なる語彙の形成が促進される，という仮

[1] 実際には完全に白紙ではなく，特定の言語（母語）に対する「認知バイアス」を持つと言われる．
[2] 単語やフレーズなどの意味上のまとまり
[3] 例えば「りんご」という単語は「り」「ん」「ご」というモーラ列となる．

説である [8, 9, 10, 11].

3.2.1　NLM-e

NLM-e の立場では，先ほども述べたようにまず音韻や音素といった知覚の単位での形成が行われるとされる．こちらの立場においても，言語獲得プロセスに対する妥当性は様々な報告によって主張されている．例えば，乳児が生まれて間もない頃にはどのような音素（例えそれが非母語の音素であったとしても）遷移に対しても等しい感度を示すが [12, 13]、生後 1 年になる頃には，母語に存在する音素の遷移に対してはより鋭敏に，そして非母語に対しては鈍感となっている事も報告されている [4]．これは，乳幼児がこの期間に周囲の音声信号から音素・音韻カテゴリを学習している可能性を示唆している．

こうした中，言語獲得に関わる/関わらないによらず、連続音声から音素・音韻を学習する試みが数多く為されてきた [3]．しかし話者の違いや前後の音による調音結合、さらに文脈や語彙に依存して音素や音韻の音響的特徴は変化することから，音声信号のみからセグメンテーションされたセグメントを正しくカテゴライズすることは不可能に近い [3]．また乳幼児に対して，音素や音韻のカテゴリについての正確な知識を仮定する事が出来ないという問題も残る．そこで NLM-e では音素や音韻の学習は語彙や視覚情報，幼児本人の調音活動などの情報を仮定した統合的なフレームワークが主となっている．

3.2.2　PRIMIR

幼児は言語能力の発達の過程で，まず始めに 1 単語から成る言葉を発するようになり，その後に助詞を用いない複数の単語から成る言葉を話す様になる．これらの様子から，単語やフレーズなどの語彙トークンが幼児語の単位であると考える事も出来る．これを裏打ちするように，幼児は音韻的意識が希薄であることと，単語を個々の音韻を抽出する能力が獲得されるのが小学校低学年あたりであるという報告も存在する [14, 15]．また音韻障害の研究では，音声コミュニケーションが問題なく行える子供達の中に音韻意識が弱く，文字の書き取りや読み取りにリスクのある集団が確認されている [16]．彼らは音声コミュニケーションが可能にも関わらず音声を音韻に分解して処理する能力が欠如しているという事実が，言語獲得における語彙形成の先行を示唆すると共に，子供達の母語習得の過程が音声コミュニケーション能力の獲得，音韻意識の発達，識字能力の発達という順で行われる事を示していると言える．実際に発達心理学の分野においても，子供達が発達の段階で「しりとり」などの言葉遊びを通じて音韻意識を育むとの報告もある [17].

また工学的な視点に立っても，音素や音韻系列の学習には語彙やコンテキストの情報が不可欠となる事から，音声信号のみからモデルをボトムアップに構築する場合には，音声を語彙トークン単位で取り扱った場合の方が都合が良い．特に語彙情報によって同一の発話内容と紐付けられた音声から音韻系列を学習するプロセスは，正に現在幅広く用いられている HMM モデルの学習と非常に類似している．

3.3　言語リズム

3.3.1　言語に固有のリズム

人が会話をするときの重要な要素の一つとして言語のリズムが挙げられる．世界中の言語はリズム上の特性から「強勢（ストレス）と強勢の間隔が一定」の強勢拍リズム,「音節と音節の間隔

が一定」の音節リズム,「モーラとモーラの間隔が一定」のモーラ（拍）リズムの 3 つに大別される [18]．非ネイティブ話者の英語発声において，このリズムの違いがネイティブ話者による音声知覚精度に影響しているという報告もされている [19]．また近年では，言語を獲得する前の乳児は母語のリズムに敏感であり，知覚したリズムを用いて母語とそれ以外の言語との弁別を行っているという報告や [20]，強勢拍リズム言語を持つ幼児がストレスの位置を手掛かりに単語境界の判別を行っているなどという発達心理学の立場における研究結果もある [21, 22]．このように言語のリズムは音声を認識したり，母語を獲得する際に重要な役割を果たしていると考えられる．言語リズムの構成要因の中でも，強勢弱勢は音節の発音に対する特徴であり，モーラは音節のサブセットであることから，少なくとも音声を知覚する際には音節のような単位を目安にしていると考えられる．近年では言語情報や音素情報などの前提知識を使用せずに音節情報を得る研究がなされており [23, 24, 25, 26, 27, 28]，そういった技術の向上はより人間らしい音声技術の発展に繋がるだろうという期待される．

3.3.2　音節と聞こえ度の関係

i)　一般的な音節 (Syllable) の定義

音節は一般的に母音が複数の子音を引き連れる構造と定義されており，この様子は母音 (vowel) を V，子音 (consonant) を C とすると C*VC* のように表すことができる [18]．また，母音のみ（V）の場合でも音節となることがある．この定義より，音節の主となる音節核は母音のことを指すことがほとんどだが，子音のみで構成される部分が音節として知覚されるような場合もある [30]．英単語の twinkle の後半部分の音節/kl/の/l/は syllabic consonant の典型である．また，CV や CCV，CCCV のように母音で音節が終わるような音節は開音節と呼ばれ，日本語の音節のほとんどは/ku/，/ha/のような開音節であることが知られている [18]．しかし，/des/のように「す」の母音が無声化されたものや/meN/のように子音である/N/で終わる音節などいくつかの例外も存在する．この場合，/des/，/maN/はそれぞれ 1 音節である．

ii)　聞こえ度 (sonority)

聞こえ度とは，音声の聞き手側が感覚する音の固有の大きさであり，明確な物理量での説明は難しく様々な定義がなされている．例えば [31] では「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」と定義されており，[18] ではさらに広い概念として「音の固有の大きさ」として説明されている．音は聞こえ度の大小によって図 3.1 のような聞こえ度の階層構造 (sonority hierarchy) を作る．図 3.1 は日本語音声の各母音・子音の聞こえ度の大きさのおおまかな分類である．母音の中でも口を大きく開ける/a/の音が最も大きく，逆に口を小さく開ける/i/や/u/では相対的に聞こえ度が低いというように，聞こえ度を物理的な人間の音声生成活動に結びつけようとする研究も一部ではなされている [34]．

iii)　聞こえ度を用いた音節の定義

音節は CV や CVC のように母音が複数の子音を前後に引き連れた構造を持つと説明したが，聞こえ度によって音節の構造を定義することもできる [18, 31, 32]．図 3.2 のように音節は聞こえ度の高い母音（音節核）を中心として前後に聞こえ度の低い音を引き連れ聞こえ度が滑らかに変化する（聞こえ度連続の原理）．また，音声は音節のシーケンスなので聞こえ度の山と谷ができることになる．よって，音声を聞こえ度で表したときにできる山のひとつひとつを音節と見なすこ

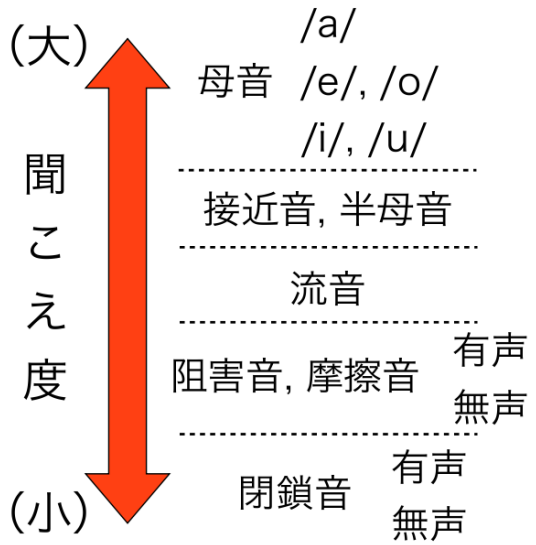

図 3.1: 音の聞こえ度

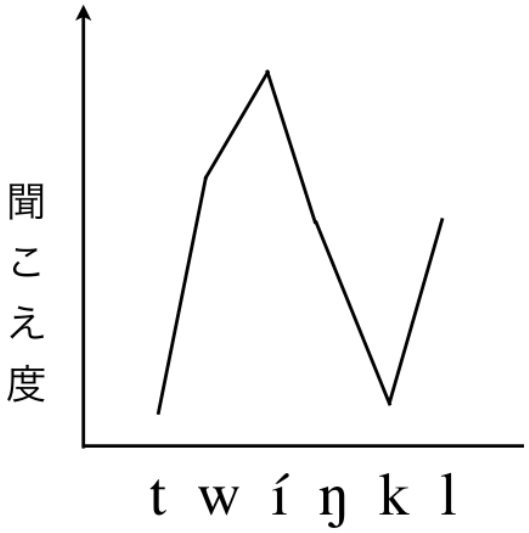

図 3.2: 聞こえ度の変化

とができて，そのピークを音節核と定義することができる．この聞こえ度の概念を用いれば前節で例外として挙げた twinkle の/kl/の場合も，子音の中では比較的聞こえ度の大きい/l/の音が母音ほどの大きさでは無いが，小さな聞こえ度の山を形作っていると説明することができる．

3.3.3　聞こえ度の物理的定義

物理量によって聞こえ度を推定する試みは，これまでにいくつかの研究で検討されている．例えば波形包絡情報を用いて音節情報（核・境界の位置）の自動推定を行うという手法はその代表的なものであり [25]，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．同様に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 [26]．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．また，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 [27]．ここ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，聞こえ度の大きさはスペクトルの安定度であると解釈している．

3.4　語ゲシュタルト

3.4.1　語ゲシュタルト

幼児が家族などの非常に少ない話者サンプルから得た音声のみから，次第に言葉を覚えて言語を得て，そして誰の声でも認識できるようになる．この過程を音声認識システムに当てはめると，非常に少ない学習データから話者性の情報などをそぎ落とす言語情報のみを取り出すことにより，入力されたサンプルを不特定多数の話者に汎化していることになる．幼児は両親や家族などの声を真似る事によって言語を獲得することになるが，ここで幼児は声のどの部分を真似ているかが

問題となる．前述した通り発達心理学の立場では語ゲシュタルトとよばれる単語全体の語形・音形をまず獲得し，その後に個々の文節音を獲得すると仮説が存在する．語ゲシュタルトは一つ一つの要素（特徴時系列）を個々に捉えるのでは無く，非言語情報とは独立な単語全体の語形・言語情報そのもの定義される [10].

3.4.2　語ゲシュタルトの物理的定義

現在の音声技術では言語情報と非言語情報を完全に切り離す事は出来ないが，乳児の言語獲得プロセスの観察などから，乳児にもこの語ゲシュタルトを直接捉える能力が備わっているとも考えられる．語ゲシュタルトの物理的解釈として，峯松らの構造的表象が挙げられる [2]．構造的表象では，音声の非言語的情報を取り除き，言語的情報のみを抽出する事を試みている．

3.5　本研究の立場

本章のここまでは発達心理学の知見に基づいて，どのようなアプローチがより乳児の言語獲得に即したものであるかを議論してきた．そのなかでも特に音韻が先か (NLM-e)，それとも語彙が先か音韻が先か (PRIMIR) という問題は，システムの設計に多大な影響を与えるので慎重に選択する必要がある．ここで言語リズムの例から乳児の知覚の単位が音節であるという事や，語ゲシュタルト（単語全体の語形・概形）を捉えているとの報告されている事から，本研究では PRIMIR の立場を取り，以降の章では (1) 如何にして言語リズムに即した音声処理やモデル化が行うのか，(2) 如何にして音声から教師無しで語彙を獲得するのか，という問題について実験・検討を行う．

第4章

言語リズムの抽出と
それに即した音声モデリング

4.1　はじめに

本章の前半部分では幼児は言語のリズムに敏感に反応するという報告 [20] に基づき，言語リズムを幼児が連続音声を捉える際の目印と考えてこれを音響特徴のみから推定する手法について検討を行う．この言語リズムとは「聞こえ度」の変化パターンとして考えることができる．そこで音響的に定義した聞こえ度パターンから音節核を抽出する手法について実験を行った．そして後半部分では音韻意識の発達を調査した研究例 [8, 16] 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．そこで話者性を（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，本研究では上記の問題を構造表象を用いて検討する．

4.2　波形包絡を用いた音節核の自動推定

4.2.1　従来手法

Rudi らの研究 [23] では，波形包絡情報を用いた音節境界の自動推定を行ういくつかの手法を紹介し，それらの手法による音節分割の精度を検証している．この論文内で紹介されている手法のひとつである，Mermelstein の極小法 [24] での音節境界の抽出の流れを説明する．まず高域強調などの前処理を施した音声をバンドパスフィルタ（BPF）に通して帯域を制限する．そしてその波形の包絡を抽出し，包絡の極小値を音節境界の候補とするという流れで音節境界候補の推定を行っている．他にも同論文内で紹介されている R．Villing の提案した手法では [25]，音声の周波数帯域をバンドパスフィルタで制限する代わりにフィルタバンク出力を用いて，最終的に抽出された音節境界を統合している．これらの手法は波形包絡が聞こえ度に対応していると考えて聞こえ度を推定している．そして，その推定した聞こえ度の谷が音節境界に相当するとしてを音節境界を抽出している．しかし，いずれの場合も音節境界を求めるというタスクの都合上，極小値をそのまま境界としているため，波形から包絡を抽出する際のスムージング周波数（ローパスフィルタのカットオフ周波数）が高くなると，本来意図していない無音区間や包絡上の小さな振幅の上下などで膨大な挿入誤り[1]が発生してしまう．論文内で 99%を越える Match Rate[2]が示されているが，結局は隙間無く挿入された境界候補が押し上げているに過ぎないことが分かる．

4.2.2　提案手法

ここでは Mermelstein の提案した方法 [24] の応用として，波形包絡を用いた音節核抽出の枠組みを提案する．図 4.1 に音節各の抽出の流れを示す．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[3]得られた波形を全波整流[4]したのちに，10 Hz から 50 Hz 程度のスムージング周波数[5]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てそれを音節核とする．

[1] 本来なら境界が存在しないところに境界を推定してしまう誤り．
[2] 全音節境界の内で推定することのできた割合，Recall と同義．
[3] ここでバンドパスフィルタの帯域を 500 Hz から 2000 Hz 前後とした理由は，人が会話するのに必要な言語情報は概ねこの範囲に収まっているとされているからである [29]．
[4] 振幅の負成分を正成分に折り返す操作のこと．
[5] 音節の継続長は言語によって異なるが概ね 20 ms から 200 ms あたりであることから，包絡を取る際のカットオフ周波数を 30 Hz から 60 Hz 程度としている．

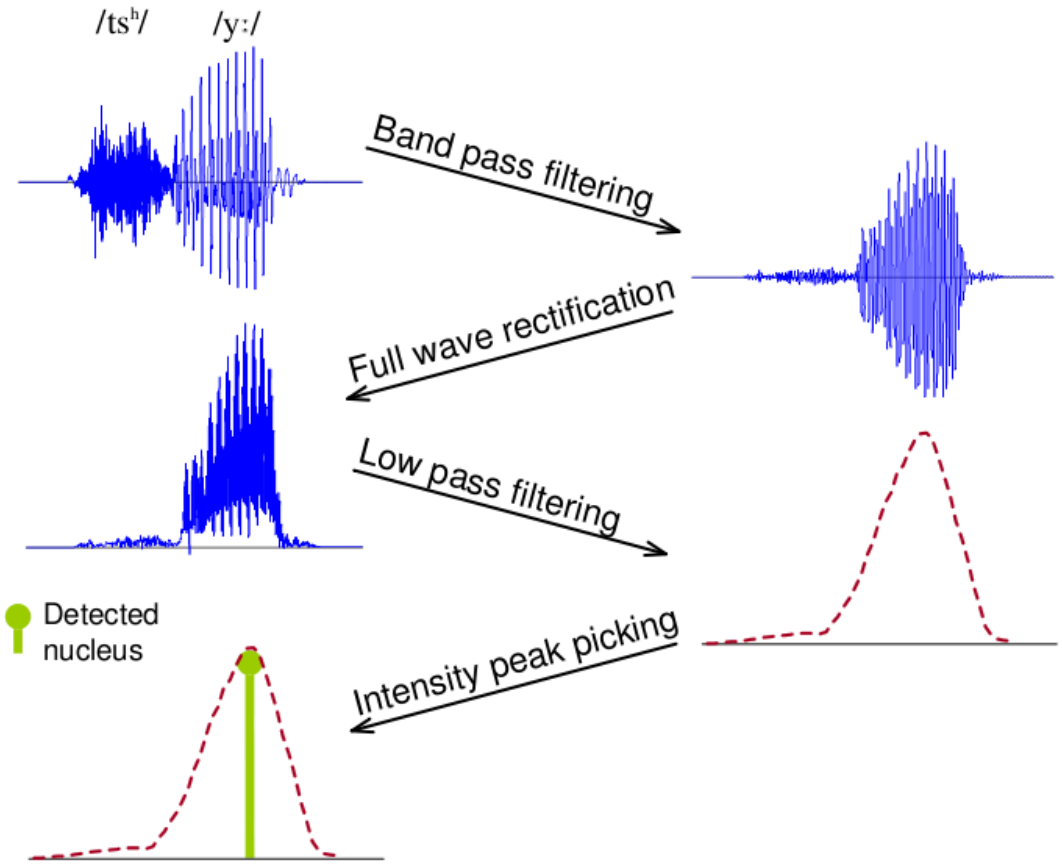

図 4.1: 波形包絡を用いた音節核抽出の流れ

従来手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，図 4.2 のように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．

また，初期検討の段階で実際に抽出された音節核を見てみると，/k/や/t/などの無声破裂音での挿入誤りが非常に多くなった．これらの子音は無声音なので波形全体のエネルギーは小さくなるが，発声の開始部分で破裂による急峻な振幅が包絡に残ってしまうことが多々あり，波形包絡や波形のパワーを用いて音節核を抽出する枠組みでの誤りの主な要因となっている．そこで，母音の無声化などの例外はあるものの，母音は基本的には有声であることから [18]，音声解析ツール

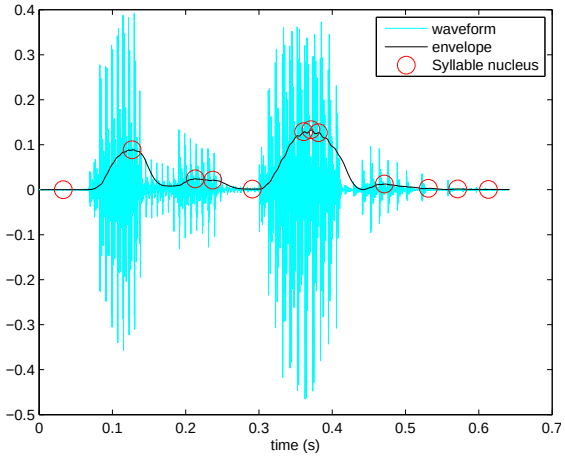

図 4.2: 従来手法の音節核のピックアップ

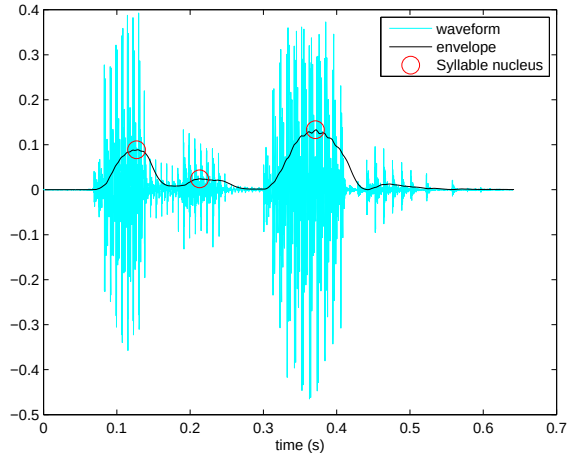

図 4.3: 提案手法の音節核のピックアップ

表 4.1: 実験条件：音節核抽出パラメータ

BPF の帯域	500 Hz–1500 Hz, 500 Hz–2000 Hz
スムージング周波数	40 Hz, 50 Hz, 60Hz
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

の praat[6]を用いて得られた基本周波数情報を参照し，無声区間から抽出された極大値をカットするという操作を追加した．これらの方法を用いて音節核を抽出した結果が図 4.3 であり，確かに上記の問題が軽減されている事が分かる..

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する事ができ，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，この所望する数の音節核を抽出できる特徴を利用する．

4.2.3　実験

i)　実験条件

実験データには「日本人学生による読み上げ英語音声データベース（ERJ）」の「文強勢、文リズムに関する文（120 文）」の英語母語話者の男性 8 人、女性 12 人によるサンプルの部分の計 1200 発声を使用した。今回の実験では，極大抽出区間を 50 ms とし，極大値を抽出する際の閾値を最大振幅の 10%として行った．バンドパスフィルタの帯域とスムージング周波数を変化させた結果を考察する．評価指標には Recall，Precision と F 値を使用した．

[6] http://www.fon.hum.uva.nl/praat/

表 4.2: 抽出精度の客観的評価

LPF	500 Hz–1500 Hz			500 Hz–2000 Hz		
BPF	再現率	適合率	F 値	再現率	適合率	F 値
40 Hz	0.735	0.805	0.768	0.743	0.804	0.772
50 Hz	0.742	0.807	0.773	0.751	0.806	0.778
60 Hz	0.747	0.806	0.775	0.756	0.808	0.781

表 4.3: 抽出精度の主観による評価

再現率	適合率	F 値
0.857	0.923	0.889

ii)　結果：客観評価

客観評価は，HTK（HMM Toolkit）[7]の強制アライメントの機能を用いて発声の音素ラベルを作成し，ラベルの母音の区間に抽出した音節核が存在すれば，それを正解とするように行う．ここで強制アライメントに必要な単語辞書は Carnegie Mellon University (CMU) の公開している cmudict0.7a[8]を，音響モデルには Wall Street Journal[9]の公開している混合数 16 のガウス混合モデルで学習された triphone モデルを用いた．実験結果を表 4.2 に示す．条件による大きな結果の違いは見られず，この枠組みでの評価結果では，Recall は 7 割半ば，Precision は 8 割前後で頭打ちになっている．

しかし この評価方法では，音節内には存在するが若干母音の位置から外れている音節核や syllabic consonant の評価が抜け落ちてしまったり，無声化や話し方による音節数のゆらぎ，そのほかにも 2 重母音や 3 重母音と母音の連続の区別をどう解釈するかなどの問題点が存在する．また，これらの要因により同じ単語でも発声ごとに抽出される音節の数が変化することになる．例えば，2 音節単語の「みかん」を発声したとしても，ある話者では 3 つもの音節核が抽出され，ある話者では 1 つの音節核しか抽出されないということも起こる．ここで用いた評価方法は簡便的な評価には有効ではあるが，これらの要因を正しく評価できていないので不十分であると言える．

iii)　結果：主観評価

前節で述べた問題を解決するため，英語母語話者の主観による抽出結果の評価を行った．音声を実際に聞いて音節を特定し，包絡から抽出した音節核が妥当であるかをひとつずつ手作業で評価する．結果の評価は米語母語話者であり言語リズムを研究されている昭和音楽大学の Donna Erickson 教授に依頼した．実際の評価は前節で得られた抽出結果から，バンドパスフィルタが 500–1500Hz でスムージング周波数が 50Hz の場合のものを使用した．また，作業の効率を考慮して，評価は文章バランスを考慮して抜き出した 210 発声に対して行った．以下に評価方針を示す．

- 音声から実際に感覚される音節核（Syllabic consonant でもよい）の位置から考えて妥当な位置に抽出された音節核が有れば，その音節核は正解とする．
- 音声から実際に感覚される音節核の位置から考えて妥当な位置に音節核が抽出されなければ，その部分は削除誤りとする．

[7] http://htk.eng.cam.ac.uk/
[8] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[9] http://www.keithv.com/software/htk/us/

- 音声から実際に感覚される音節核の位置から考えて妥当では無い位置に抽出された音節核が有れば，そのピークは挿入誤りとする．
- 音節の定義に捉われず，実際に音声を聴き，感覚される音節核情報を基に評価を行う．

主観評価結果を表 4.3 に示す．こちらの評価では一人の主観に大きく作用されるものの，上記の問題が解消されていることに加えて，より聞き手の感覚に即した評価となっている．

iv)　考察

どちらの評価方法でも Recall，Precision 共に比較的良好な精度が得られているが，特に主観評価の結果が非常に高くなっている．このような結果は，幼児による言語獲得を模擬するシステムの構築という本研究の目的に非常に良く合致していると言える．本手法のみでは連続音声から音節へのセグメンテーションを行う事が出来ないが，音節核を中心に大枠で音韻的なまとまりを捉えそれらを何らかの手法を用いて照合するような枠組みでは，この音節核位置の情報はアンカーポイントとして非常に強力な情報となる．また，こういった言語リズム情報は音声のセグメンテーションのみならず，音声のモデリングに応用する事も考えられる．

4.3　音節核情報を用いた構造的表象による孤立単語認識

4.3.1　従来手法：音声の構造的表象

i)　音声に含まれる情報

音声に含まれる情報は言語情報と非言語情報，パラ言語情報の 3 種に分類される．言語情報とは発声内容などの情報であり，非言語情報は話者による声道の長さや形状の違い，雑音などの情報，そしてパラ言語情報は韻律などによる話者の意図や感情の情報である．現在主流となっている音声認識の枠組みでは，音響モデルにおいて，HMM（隠れマルコフモデル）を用いて話者性の違いなどの非言語情報を隠れ変数として扱うことにより発声における言語的情報のモデル化を行っている．しかし，このモデルは話者性が大きく異なる[10]入力に対しては頑健に適用できないことも多い．このようなアプローチには必ず言語情報以外の情報も含まれてしまい，結局は言語情報のみをモデル化しているとは言い難い．

ii)　音声の非言語的特徴のモデル化

上述した通り，音声には言語的特徴の他にも非言語的特徴が不可避に混入されることが分かっている．音声に混入する非言語的情報を大別すると，主に加算性歪み，乗算性歪み，線形変換性歪みの 3 種類になる [2, 33]．加算性歪みとは音声のスペクトル領域での加算で表現され，例として背景雑音などが挙げられる．こうした雑音は，防音室で録音するなどすれば取り除ける．スペクトル領域への線形変換性歪み（$\boldsymbol{A}$）と乗算性歪み（$\boldsymbol{b}$）の影響を図 4.4 に示す．乗算性歪みは音声のスペクトル領域における乗算で表現される歪みであり，マイク特性などが例として挙げられる．スペクトル領域における乗算は，対数パワースペクトルに変換する過程で加算に変換されるので，ケプストラムを $\boldsymbol{c}$ とすると，非言語情報によって歪められたケプストラム $\boldsymbol{c}'$ は $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{b}$ のように表される．また，線形変換性歪みとは話者の声道長や声道形状の差異が原因となる歪みであり，図 4.4 の $\boldsymbol{A}$ で表されるように，フォルマント周波数が周波数軸方向に移動することを表

[10] 例えば身長 2m の巨人や身長 70cm の小人の声道長の違いは音響的にも大きな違いとなる．

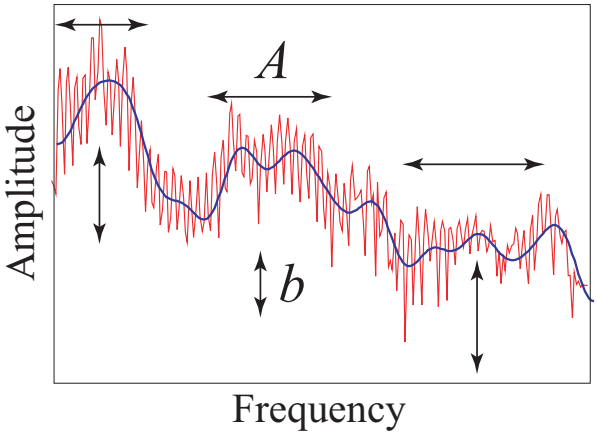

図 4.4: スペクトルに対する線形変換性歪み（$\boldsymbol{A}$）と乗算性歪み（$\boldsymbol{b}$）

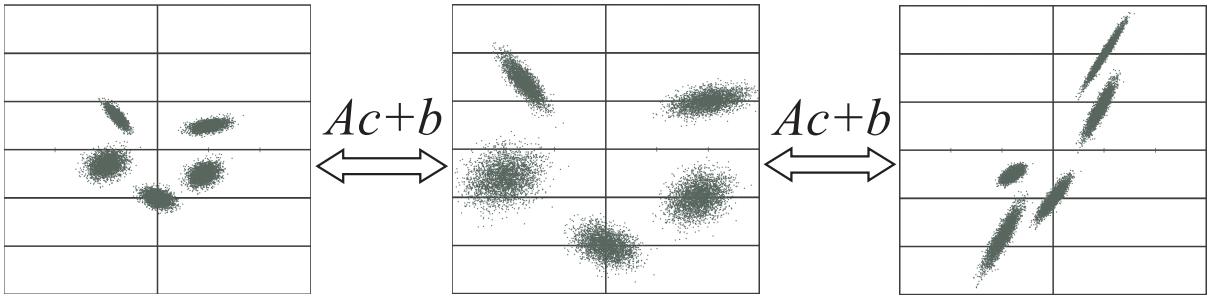

図 4.5: アフィン変換による分布群の変化（これらは全て同一の構造を持つ）

す．この線形性変換性歪みによる影響は $\boldsymbol{c'} = \boldsymbol{Ac}$ という式で近似されることが分かっている [35]．よってこれらをまとめると，音声の音響的特徴量に不可避に混入する非言語特徴量による歪みはケプストラムに対するアフィン変換 $\boldsymbol{c'} = \boldsymbol{Ac} + \boldsymbol{b}$ で表現されることになる．これにより，音声の非言語特徴量による歪みをケプストラムに対するアフィン変換による空間写像でモデル化することができた．

4.3.2　音響的普遍構造

前節でケプストラムに対する歪みをアフィン変換でモデル化した．これらの歪みの混入は不可避であるので，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は分布間距離である f-divergence が挙げられる [2]．ある分布 p_1 と p_2 の間の f-divergence は以下の式で表される．

$$f_{div}(p_1, p_2) = \oint p_2(x) g\left(\frac{p_1(x)}{p_2(x)}\right) dx \tag{4.1}$$

実際の実験では，f-divergence の式において $g(x) = \sqrt{x}$, $BD(p_1, p_2) = -\ln\ f_{div}(p_1, p_2)$ として表されるバタチャリヤ距離を用いる．分布 p_1 と p_2 の間のバタチャリヤ距離は以下のように表される．

$$BD(p_1, p_2) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x) p_2(x)} dx \tag{4.2}$$

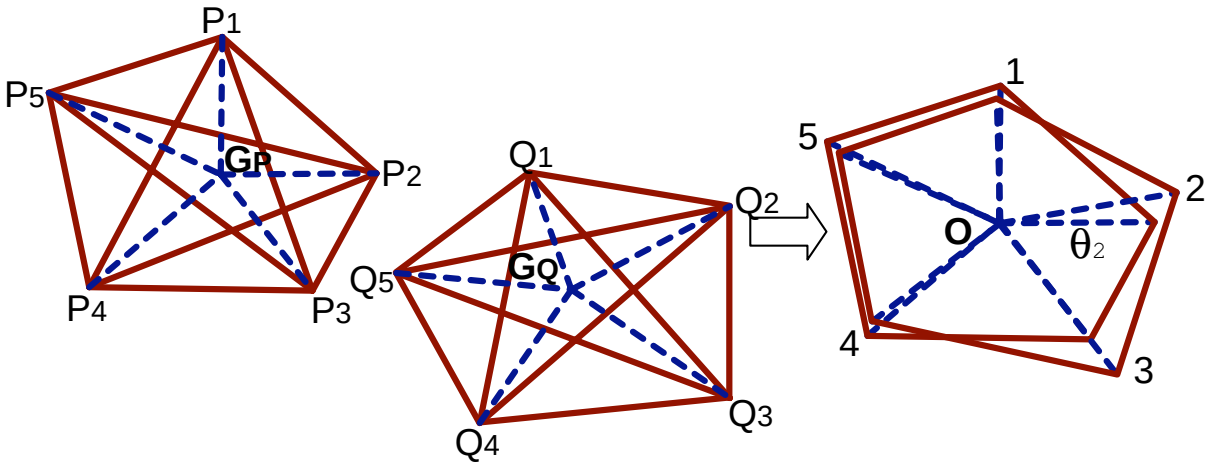

図 4.6: 回転及び平行移動を通して行なう構造照合

分布 p_1，p_2 をそれぞれガウス分布であるとすると，バタチャリヤ距離は以下の式のように簡便な形に展開できる．

$$BD(p_1,p_2)=\frac{1}{8}\mu_{12}^{T}V_{12}^{-1}\mu_{12}+\frac{1}{2}\ln\frac{|V_{12}|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}} \tag{4.3}$$

ただし，$p_1=\mathcal{N}_1(\mu_1,\Sigma_1)$，$p_2=\mathcal{N}_2(\mu_2,\Sigma_2)$ であり，$\mu_{12}=\mu_1-\mu_2$，$V_{12}=\frac{\Sigma_1+\Sigma_2}{2}$ とする．ケプストラムベクトルの系列を N 個のガウス分布に割り当てて表現されるとすると，${}_NC_2$ 個のバタチャリヤ距離によって張られる距離行列が得られる．バタチャリヤ距離は前述した通り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，これを音響的普遍構造と呼ぶ．図 4.5 に，アフィン変換による分布群の変化を示す．これらは一見まったく異なる分布群に見えるが，アフィン変換によって変換されたものならば，その分布間距離は不変となっている．

4.3.3　構造に基づく音響的照合

前節までで求めた音響的普遍構造の音響的照合について説明する．構造間の類似度は，一般的には図 4.6 のように M 個の頂点で構成される二つの構造 $(P_1,P_2,...,P_M),(Q_1,Q_2,...,Q_M)$ のうち，一方をシフト（b）と回転（A）のみにより他方に近づけたときの，頂点間の距離の和として求められる．

ここで，距離行列を求める際にバタチャリヤ距離の平方根を用いると，上記のようにして得られる P_{ij} と Q_{ij} の構造間距離が，以下の式で近似されることが分かっている [33]．

$$\sqrt{\frac{1}{M}\sum_{i<j}(p_{ij}-q_{ij})^2} \tag{4.4}$$

ただし，$p_{ij}=\overline{P_iP_j},q_{ij}=\overline{Q_iQ_j}$ であり，p_{ij},q_{ij} はそれぞれ構造 P，Q を表す距離行列 ij 要素である．この性質により明示的に構造の変換を行うことなく，音響的照合を行うことが可能となっている．

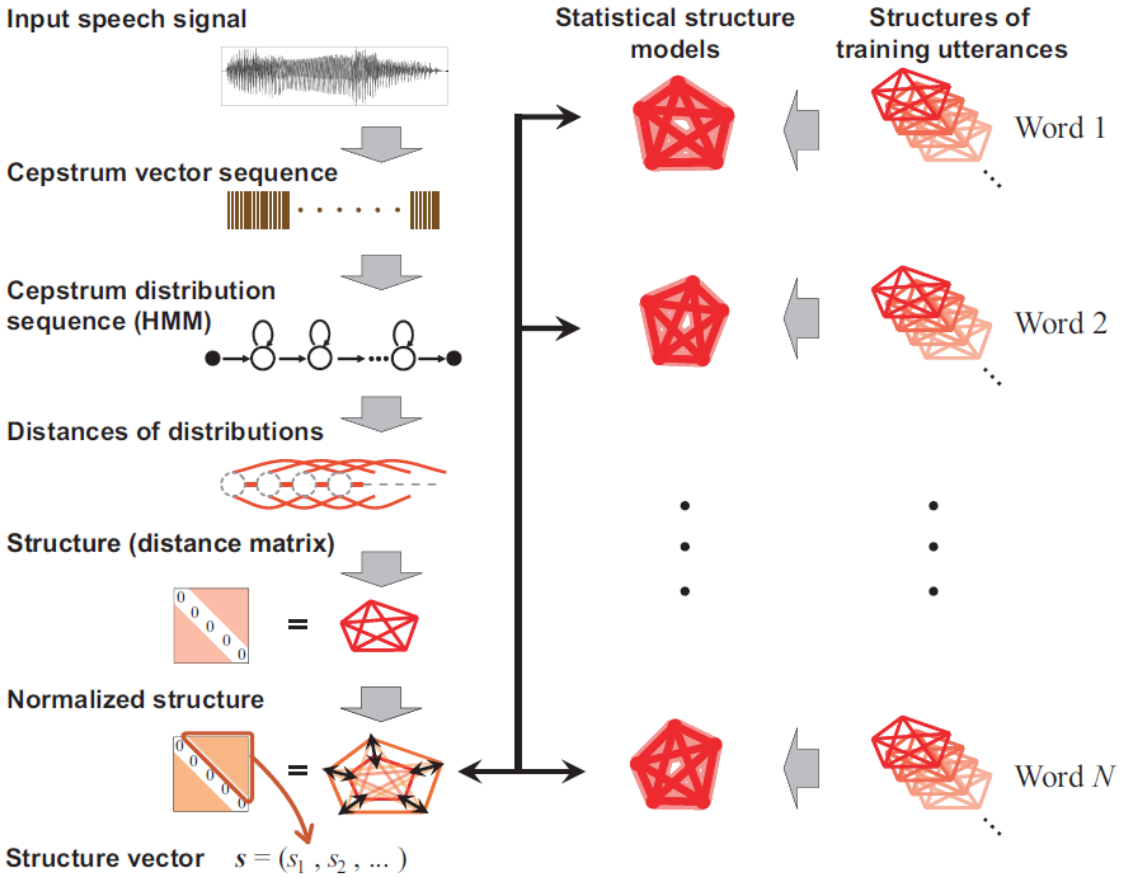

図 4.7: 構造的表象に基づく孤立単語認識の枠組み

4.3.4　構造的表象に基づく音声認識の枠組み [2]

一連の音声の構造的表象に基づく音声認識の枠組みを図 4.7 に示す．まず音声波形をケプストラムベクトルの系列に変換して，得られたケプストラムベクトルの系列からケプストラム分布の系列を推定する．ここでケプストラム系列の分布化は，最尤（Maximum Likelihood; ML）推定などを用いて HMM の各状態における出力確率分布を推定することにより行う．この HMM の各状態がケプストラムベクトルの分布系列にほかならないので，この各状態間のバタチャリヤ距離の平方根を計算することにより，分布間の距離行列を得る．この距離行列は対角成分を軸として対称となるので，この対角成分を除いた上三角成分を順に取り出して並べたものを構造ベクトルと呼び，これを認識に用いる．最終的な認識としては，音声認識システムが持つ音響モデルとの照合を行う．各単語につき複数の学習データからあらかじめ構造統計モデルを計算しておき，入力構造ベクトルがその構造統計モデルから出力されるときの対数尤度を計算することによって照合を行う．この構造統計モデルは，学習データから計算された構造ベクトルを単語ごとにまとめ，ガウス分布で表現することで得られる．

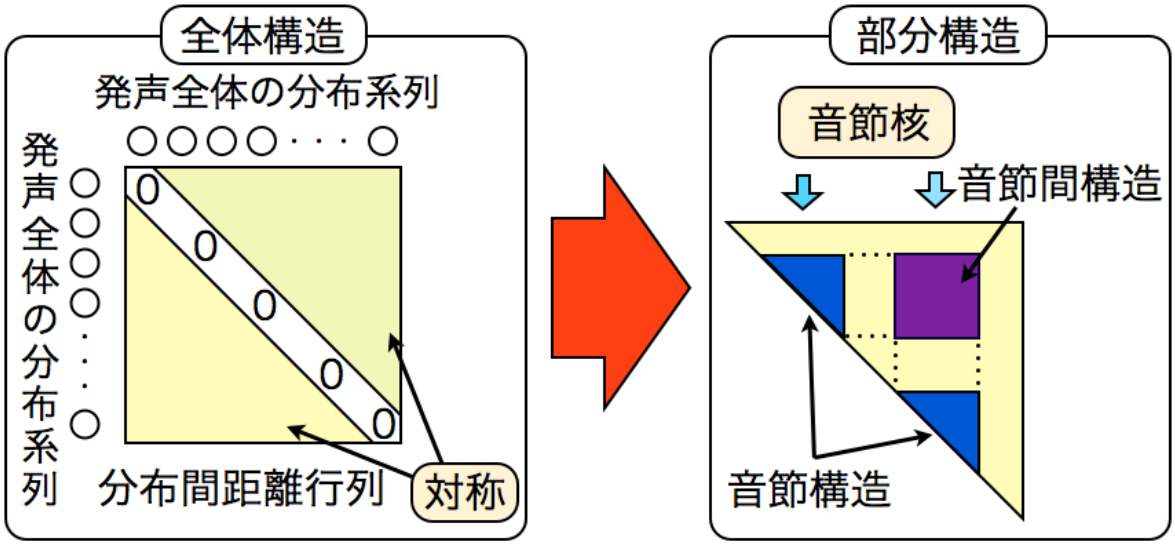

図 4.8: 部分構造の定義

4.3.5　時間アライメントの問題

音声の構造的表象を用いた音声認識の枠組みの問題として，構造の頂点間の対応関係が上手く取れなくなってしまうという，時間アライメントの問題が挙げられる [36]．例えば「あいうえお」という発声を 2 話者が行ったものを比較する際に，6 番目の状態（分布）が，一方の話者では「い」の終わり部分を表しており，もう一方の話者では「う」の始まりの部分を表しているとすると，その状態を含む分布間距離はそれぞれまったく異なる意味を持つことになり，認識精度の悪化につながる恐れがある．そこで，この頂点間の対応関係を上手く取って認識を行おうとする試みもいくつかなされてきた [36]．

高澤らの研究 [36] では，如何にこの時間アライメントの問題を解決するかを検討していて，動的計画法や制限付き HMM を用いて様々な実験・考察を行っている。しかし，問題の解決には至らず，母音のみの無意味語を用いた基礎的な研究に留まっている．音声の構造的表象 [2] 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．

4.3.6　提案手法

音声の構造的表象は話者性の変化に頑健な手法であり，単語全体を構造的にモデリングする手法として本研究との整合性が非常に高い．しかし DTW や HMM などを用いた手法とは異なり，この手法を用いて音声をモデリングする際に明示的な時間合わせが行われないので，特徴と言語内容のアライメントがズレてしまう時間アライメントの問題が存在する．この時間アライメントの問題により，同じ言語内容の音声をモデリングしたとしても発声によって構造が大きく異なる場合が考えられ，モデリング精度の悪化に繋がる．そこでこの時間アライメントの問題に対して言語リズム情報を応用し，明示的な時間合わせを組み込む手法を提案する．

i)　部分構造（音節構造，音節間構造）

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図 4.8 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．図 4.8 からも分かる様に部分構造は，音節核を目印に全体構造を組み直したものである．また音節構造は音節単位での構造的モデリングと見る事も出来る．その場合は音節間構造は単語内での音節の前後の連なりの関係を表している．

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．

ii)　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．

$$V = S + \omega \left(\frac{1}{N} \sum_{n=1}^{N} T_n + \frac{1}{{}_NC_2} \sum_{m=1}^{M} U_m \right) \tag{4.5}$$

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．

4.3.7　実験

i)　実験条件

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 4.4，Table 4.5 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 [38] 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．

表 4.4: 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 [37]
出力分布	対角共分散ガウス分布
状態数	20

表 4.5: 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

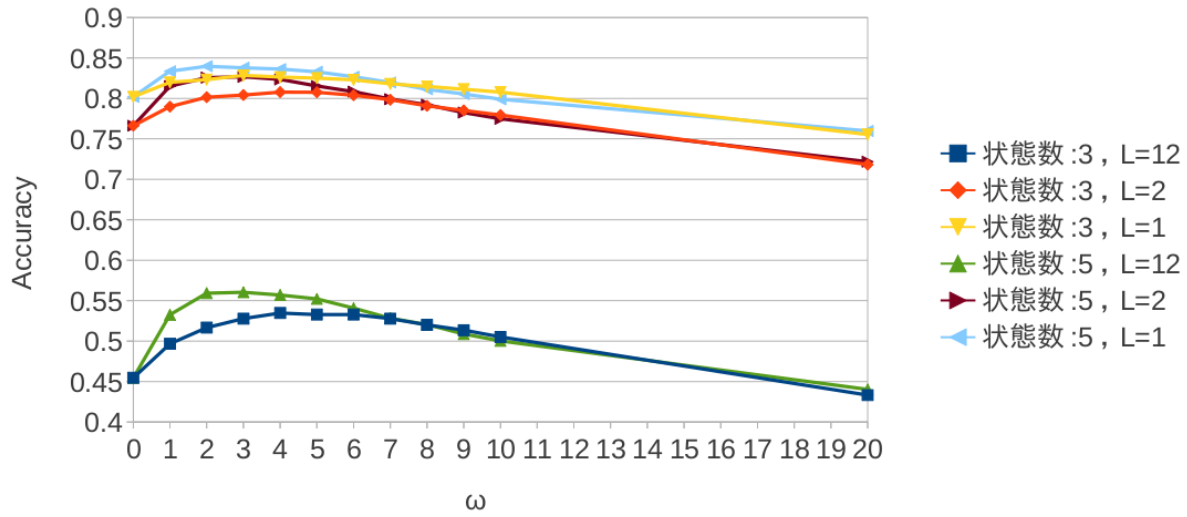

図 4.9: 孤立単語認識実験の結果

ii)　実験結果

認識実験の結果を図 4.9 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 $\omega = 0$ 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 $\omega = 0$ 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，$L = 1$，$\omega = 2$ の時の 84.0%で，ω=0 の場合と比べて誤り削減率は 19.1%となっている．

iii)　考察

当初は明示的な時間合わせを行った部分構造の重みが大きくなるにつれて認識結果が改善される事を予想していたが，実際には一定の値を超えると部分構造の重みが増えるに従って認識率は

下がり，最終的に部分構造のみを用いて行った実験ではいずれの条件でも全体構造のみでの結果を大きく下回る結果となった．このことから時間アライメントの問題は逆に，認識の際には幾らか有利にはたらいている事が示唆される．音声と統計モデルを照合する際に，時間構造が大きく異なるモデルの尤度が相対的に下がることになるので，時間アライメントの問題が一種の足切りとして有効に働いていると考えられる．ただし適切な 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，全体構造を用いた認識結果に対するリスコアリングとして言語リズム情報を用いた音節単位でのモデリングが有効であると言える．また，これらとは別に音節単位でのモデリング精度も理由の１つとして挙げられる．本実験では発声の分布系列から音節核を中心に一定幅で音節を定義したが，発声によって音節の裾野は柔軟に変化して分布系列化されている事が予想される．このことからも孤立音節を構造的にモデリングする手法と比べると，音節単位でのモデリング精度の劣化が考えられる．

4.4　まとめ

言語リズムに敏感な幼児の特性を考慮し，音響的特徴量から言語リズムを抽出するアルゴリズムの実装と，それらを用いた構造的な単語音声のモデリングを試みた．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．

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，主観評価で再現率 85.7 %，適合率 92.3 %，F 値 88.9 %と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．後半ではこの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，言語リズムを組み込んだ構造的モデリングの有効性を示す事ができた．

第5章

音声の構造的表象による頑健な教師無しパターン獲得

5.1　はじめに

前章では言語獲得プロセスを模擬するシステムの 3 つの特徴のうち,

- 言語リズムに即した音声処理を行っている.
- 単語全体の語形である語ゲシュタルトを捉えており, 非言語情報の変化に頑健である.

の 2 つを満たす音声モデリング手法について実験を行った．本章では従来の教師無しパターン発見アルゴリズムに，語ゲシュタルトの考えを導入し,

- 連続音声中から言語や音韻などの事前知識を用いずにボトムアップに語彙を獲得する.
- 単語全体の語形である語ゲシュタルトを捉えており, 非言語情報の変化に頑健である.

の 2 つの要件を満たす言語獲得システムの構築を目指す.

5.2　従来手法

5.2.1　S-DTW を用いた教師無しパターン発見 [39]

教師無しパターン発見の先行研究で代表的なものとして Segmental Dynamic Time Warping (以下 S-DTW) を用いたアルゴリズムが挙げられる [39, 40]．これは DTW ベースでパターン間類似性を計算して収集し，それを基にパターン発見を行うものである．このアルゴリズムでは大きく分けて以下の 3 つのステップにより音声信号に頻出するパターン/キーワードの発見を試みている.

1. S-DTW を用いた類似パターン候補の発見
2. 発見されたパターン候補の始終端の決定（得られた音声区間をノードと呼ぶ）
3. ノードクラスタリングに基づく語（クラスタ）の同定

ここで距離尺度としてケプストラムのユークリッド距離を採用している為，発話交替のある複数話者による音声信号を入力とすると，同じ発話内容にも関わらず複数のクラスタが形成されるという問題が発生する（教師無しパターン発見における複数話者問題）[39, 40].

これを解決するために，Glass らによって入力特徴として Gaussian Posteriogram (GP) や Universal Phone Posterior (UPP) を用いた拡張が行われている [42, 43]．これらの研究では話者情報の大きく乗ったケプストラム領域ではなく，事後確率空間で比較を行う事により上記の問題の解決を図っている．しかし，多量のデータからバックグラウンドモデルを作成しておく必要があり，本研究のように予め事前知識を仮定できないタスクにおいては使用するのは難しい.

5.2.2　SSM に基づくノードクラスタリング [41]

この問題に対して，ベクトル時系列中の任意の二時刻間のベクトル間距離により構成される自己類似度行列 Self-Similarity-Matirx (以下 SSM) に変換して比較を行うといった手法が提案されており，その有効性が示されている [41] Glass らは特徴量の段階で話者情報を消すというアプローチを行っていたが，Muscariello らはノード間距離尺度として音声特徴の相対量を用いる事で，比較的頑健なクラスタリングを実現している.

ここで使用されている SSM という特徴は，ケプストラム時系列の任意の二時刻間のフレーム距離群であり（図 5.1），こうして得られる行列は音声の相対量に基づくものなので話者の違いによ

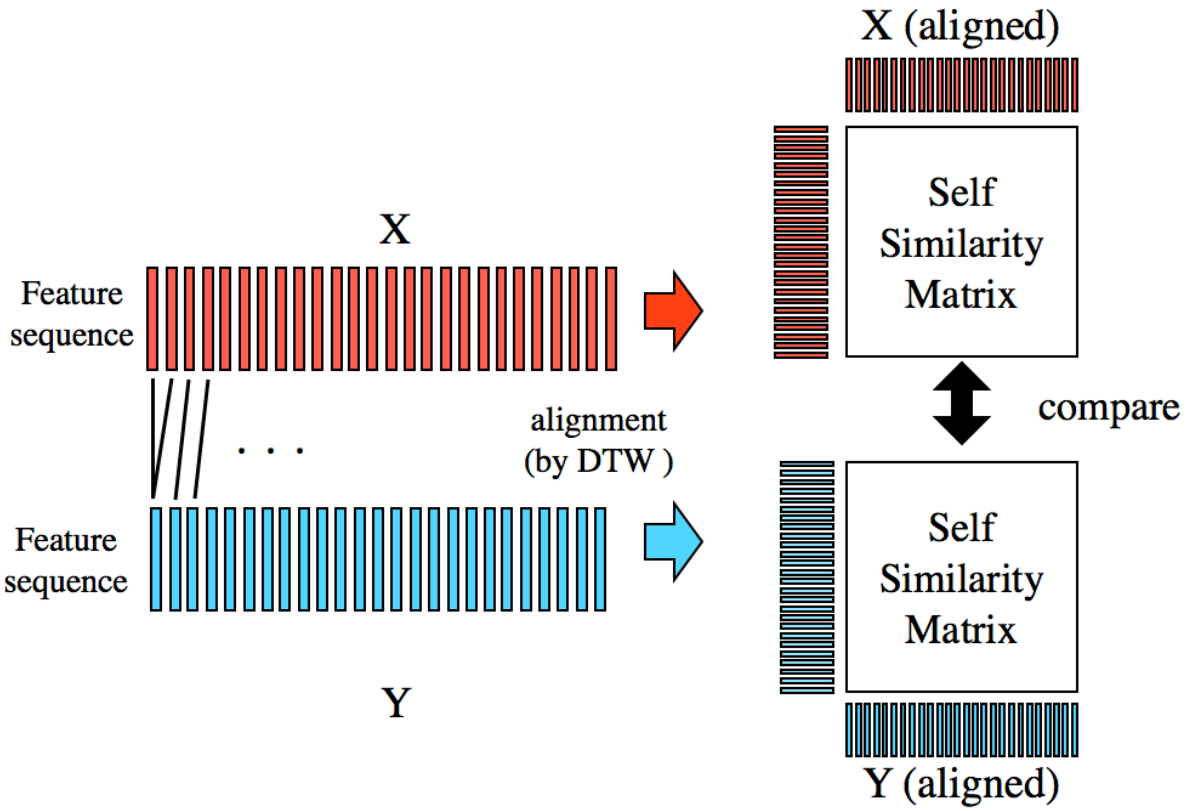

図 5.1: SSM による音響照合.

る影響が比較的少ない．この SSM 同士を比較した値をパターン間の距離尺度として用いる事で，より頑健性が高いクラスタリングを実現している．

5.2.3　音声の構造的表象による話者の変化に頑健なパターン表現 [2]

音響特徴の絶対量ではなく相対量を話者の変化に頑健な特徴として採用する SSM の考え方は，峯松らの音声の構造的表象に非常に近い [2]．音声の構造的表象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．SSM は系列間の距離行列をフレーム単位で計算しているのに対して，構造的表象では特徴量系列を一旦分布系列に変換して，それらの分布間の距離行列を特徴としている．第 4 章でも示した通り，この分布間距離に f-divergence を用いると任意の連続かつ可逆な変換に対して不変であることが数学的に保証されている．そこで本研究では従来のパターン発見アルゴリズムへの構造的表象の適用によりパターン発見の高精度化，特に複数話者音声の問題への対応を目指す．

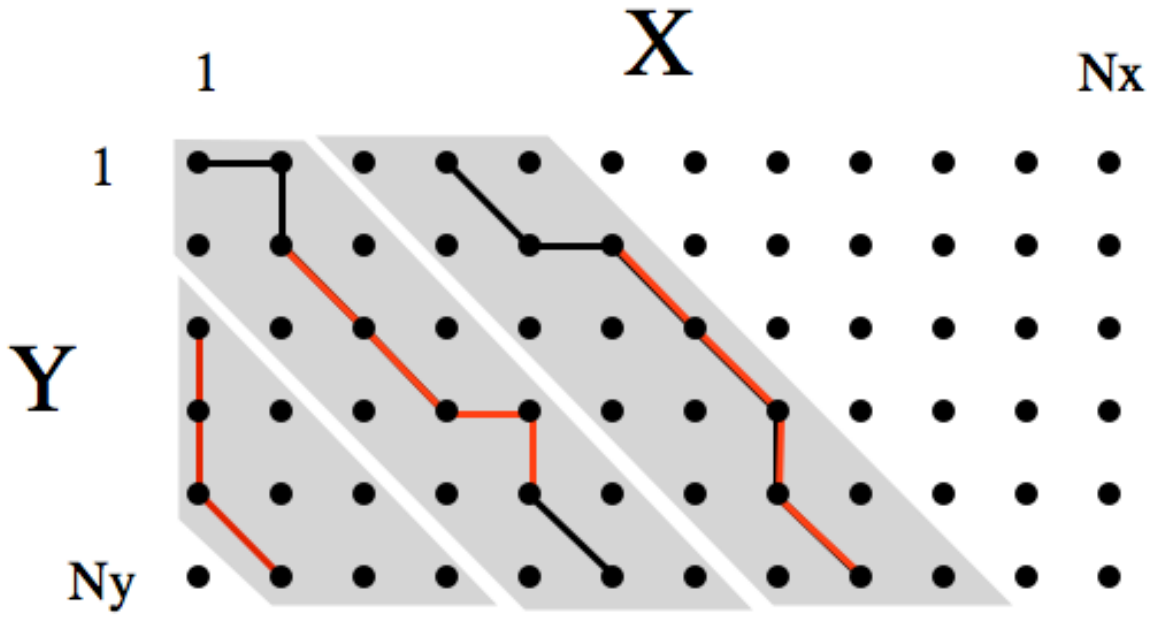

図 5.2: Segmenatl DTW と局所アライメント（赤線部分）

5.3　提案手法

5.3.1　概要

従来手法 [39] によって得られたノード情報を基に，新たなノード間距離尺度として構造的表象を導入し，複数話者データに対する精度の向上を試みる．そのため，提案手法の流れは以下のようになる．

1. S-DTW を用いた類似パターン候補の発見
2. 発見されたパターン候補の始終端の決定
3. 構造的表象を距離基準としたノードクラスタリングに基づく語（クラスタ）の同定

以下ではこれらの実装について詳しく説明する．

5.3.2　S-DTW

DTW とは時間方向に伸縮する二つの系列間の距離を計算する手法であり，同時に二系列の時間的対応も得られるので，現在では時系列間のアライメントやゲノム解析のシンボル間のパターン検出などにも用いられている．S-DTW もその一種であり，長時間の音声信号の中から類似するパターン対を見つけ出す事が目的となる．

S-DTW ではまず，Fig. 5.2 のように発話ペアから得られる距離行列をいくつかの領域（セグメント）に分割し，それぞれの領域で DTW によるアライメントを行う[1]．そうして得られた各領域のアライメントから，一定以上の長さを持ち且つ平均距離が最小となる局所アライメントを抽出

[1]基本的に処理は発話単位であり，入力が長時間の連続音声の場合は `Voice Activity Detection` などの前処理により発話単位に区切られている事が前提となっている．

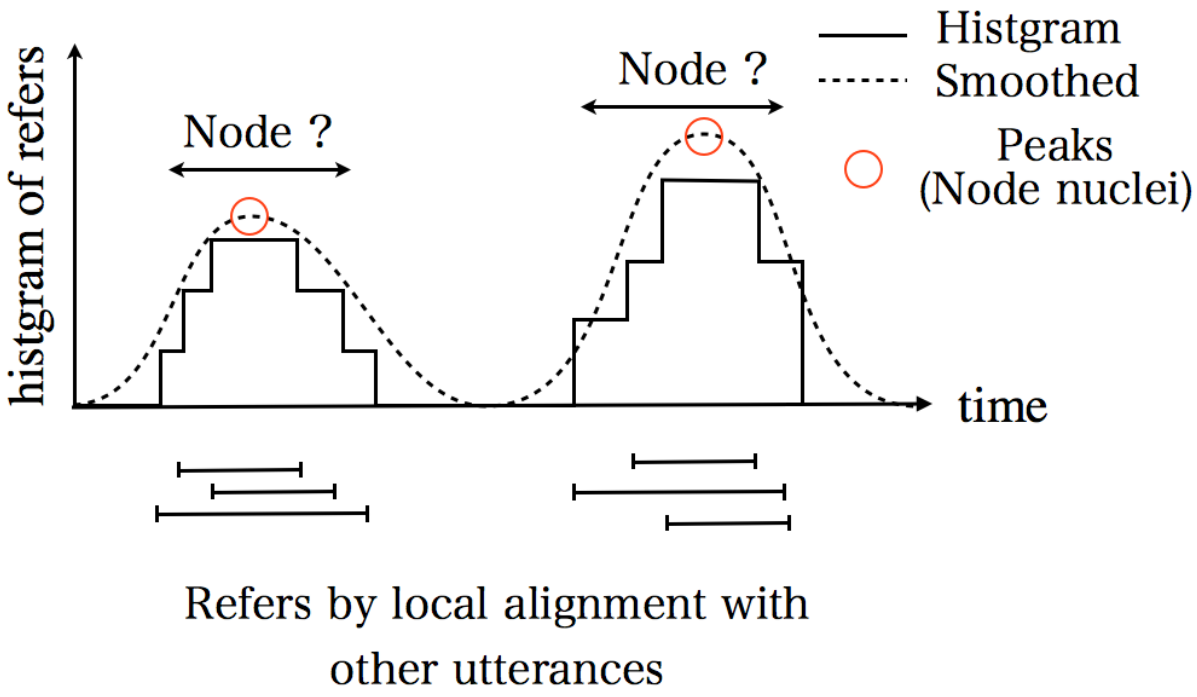

図 5.3: 局所アライメント情報による発話からのノードの抽出．発話毎に参照のヒストグラムを作成し，ノードの位置と始点・終点を確定する．

する (Fig. 5.2)．この局所アライメントは「音声信号の中のこの部分とこの部分が似ている」という情報となる．特に平均歪みが比較的小さい局所アライメントについてはパターン対が同じ単語であるか，あるいは発音的に非常に似ているという事であり，そのような局所アライメントによって何度も繰り返し参照[2]されている区間は，単語やそれに準ずる音声パターンである可能性が高い．

5.3.3　ノードの決定

S-DTW で得られる局所アライメントにはパターンの始点・終点の時刻情報が含まれるが，ペアの選び方によって区間情報が異なる事が多く，場合によってはアライメントが発話内の全く異なる単語を参照している事も考えられる．そこで従来手法 [39] に従って，この参照情報を用いて発話からノードを抽出する．参照情報からノード決定までの流れは以下のようになる．

1. 平均歪みが閾値 (θ) 以上の参照を削除
2. 各時刻毎に参照回数をヒストグラム化[3]（Fig. 5.3）
3. 得られたヒストグラムの平滑化を行い，参照数が一定以上のピークをノードの重心としてピックアップ
4. 重心を含む参照の平均開始時間と平均終了時間をノードの区間とする

こうして得られたノードは入力音声中に頻出するパターンであると考えられる．

[2]参照とは本論文では，ある発声のある区間と別発声のある区間とが類似性により対応付けられていることを言う．
[3]実際は単なるヒストグラムではなく，参照の平均歪みで重み付けを行っている．

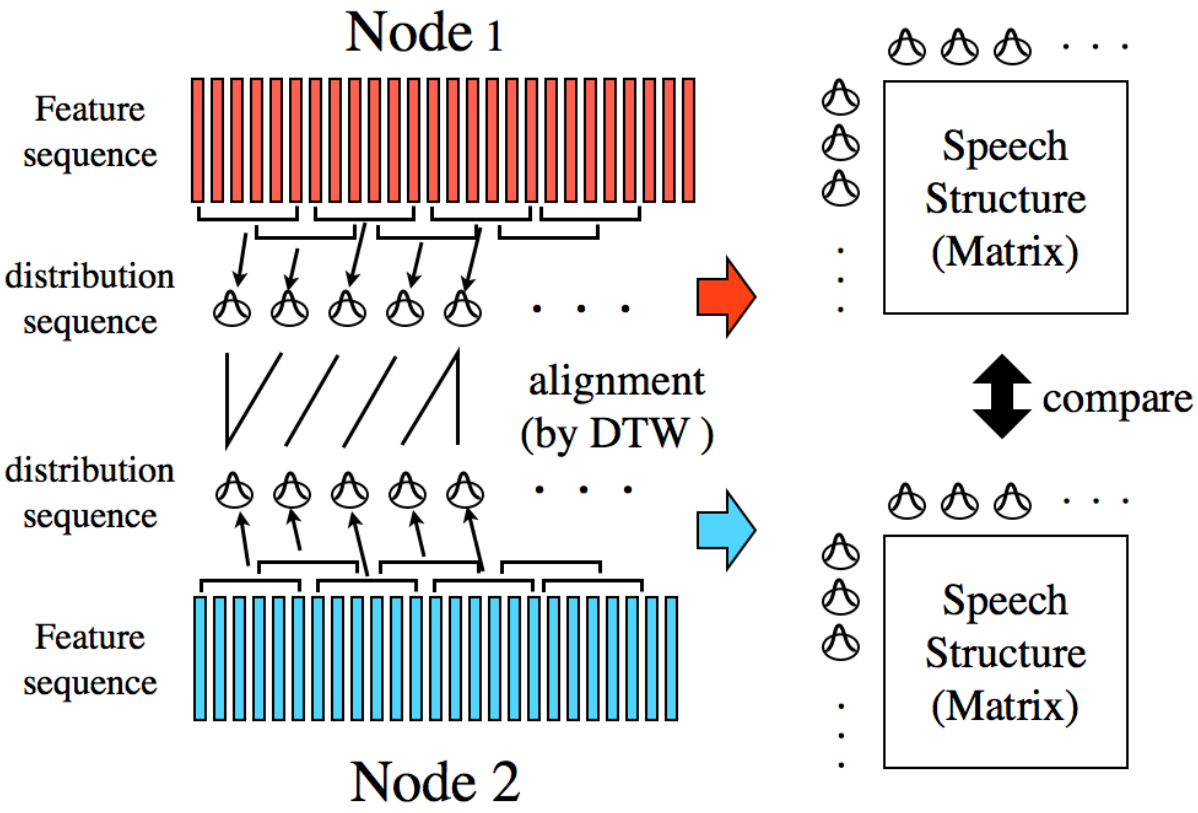

図 5.4: 構造的表象の実装

5.3.4　構造特徴を用いたグラフクラスタリング

前項で得られたノードに対して，それぞれの距離情報を基にクラスタリングを行う．ノード間距離尺度として構造的表象を用いる事で，より頑健なクラスタリングが期待できる．

構造的表象の実装はいくつかあるが，分布系列となったサンプル間の時間的対応が必要という制約が存在する．そこで SSM の実装と同じように系列間のアライメントを予め行い，サイズ合わせを行った上で比較する．特徴時系列から構造的表象を抽出し比較するまでの流れをまとめると Fig. 5.4 のようになる．ここでアライメントには分布の平均の値を用いており，構造間の比較にはユークリッド距離を距離行列のサイズで正規化したものを用いる．

5.4　実験

教師無しパターン発見アルゴリズムを複数話者タスクにおいて適用し，その結果を樹形図にして考察する．実験ではノードの抽出まではそれぞれの話者で行った上で，全話者に対して検出された全ノード間の距離を計測してクラスタリングを行う．

表 5.1: 音響分析条件

ビットレート / サンプリング周波数	16 bit / 8 kHz
窓幅&シフト長	25 ms length / 10 ms shift
特徴量	MFCC (13 dim.) + Δ + ΔΔ → 主成分分析で 10 次元に圧縮
正規化	分散：音声全体

表 5.2: 実験パラメータ

S-DTW	探索幅	700 [ms]
S-DTW	局所アライメントの最小長	80
ノード抽出	局所アライメントの閾値	上位 10%に調整
ノード抽出	参照数の閾値	3
構造的表象	分布推定の分析幅	前後 3 フレーム（計 7 フレーム）
構造的表象	分布推定のシフト長	2
構造的表象	マルチストリーム [38]	ブロックサイズ 1

表 5.3: クラスタリング結果の分析

	有効ノード数	クラスタ数	純度	エントロピー
DTW	205	32	0.93	1.19
SSM	198	21	0.72	1.21
構造的表象	125 10	0.57	0.85	

5.4.1　データベース

実験データベースには日本語連続数字読み上げコーパス AURORA2J を用いた．語彙数は 11 あり, 全て一桁の数字である[4]．話者は男性話者 MBD，MAL と女性話者 FNG，FAC の合計 4 名を使用する．

5.4.2　実験条件

実験に用いた音響特徴量の分析条件とパラメータを Table 5.1 と Table 5.2 に示す．ノードのクラスタリング手法には Newman 法 [44] を用いており，予めエッジは閾値を用いて上位 0.5%まで枝刈りを行った．また，ここではノード数が 3 以上のクラスタのみを扱う．今回は以下の 3 種類のノード間の距離尺度について検討を行った．

- 従来手法 1：DTW + ケプストラム歪み [39]
- 従来手法 2：SSM + ユークリッド距離 [41]
- 提案手法：構造的表象 + ユークリッド距離

5.4.3　結果と考察

クラスタリング結果をそれぞれ図で示す (Fig. 5.5)．図からも分かる通り，DTW ベースの距離の場合は，発話内容よりも話者の違いに敏感にクラスタリングされている事が分かる．その一方でクラスタリングの距離基準に音響相対量を用いた SSM と構造的表象の場合には，話者の違いを超えて適切にクラスタリングが行えている部分も見受けられる．どちらの場合も音韻的に近い/ichi/

[4] /ichi/, /ni/, /saN/, /yoN/, /go/, /roku/, /nana/, /hachi/, /kyuH/, /zero/, /maru/の 11 種類.

と/hachi/が話者の違いの影響を超えて一部のクラスタに集約されている様子が分かる．しかし従来手法である SSM と提案手法を比較した場合には，提案手法の優位性を論じる事は難しい．というのもクラスタリングの結果はハイパーパラメータに強く依存し，それぞれの手法における最適解を見つける必要があるからである．また，複数話者の問題は軽減する事ができたが，全体的なクラスタリングの純度はベースラインの DTW を用いた場合が高かった．このあたりはどこまでクラスタをマージして行くかという点で，エントロピーとトレードオフになっていると考えられるので，単体での性能評価の判断が難しい．

また，今回は構造的表象と SSM のクラスタリング傾向が似通ったものになってしまった原因として，前処理の DTW による位置合わせの影響が考えられる．これらの特徴を用いて比較する際に音響絶対量によるアライメントが必要となる事がボトルネックになっている可能性がある．特に現在のタスクでは前段から得られるノードの単語の一部分のみを捉えていたり，前後の単語の一部分と連結しているという事が多く，不適切なアライメント結果が得られることが多くなる．

5.5　まとめ

ベースラインとして Park らの S-DTW アルゴリズムを構築し，それに話者の変化に頑健な構造的表象を組み込む事により，複数話者音声データに対する教師無しパターン発見の精度向上を試みた．一部のクラスタでは効果が見られたが，評価基準の難しさもあり，従来手法である SSM に対して構造的表象が明確に優位であるとは言えない結果となった．その原因の一つとしてノード区間推定の精度が不安定である事が挙げられる．解決策としては細かなパラメータチューニングなどの他に，大語彙音声への移行が考えられる．今回の実験では連続数字読み上げ音声を用いたが，語彙数が極端に少ないため出現する全ての語が頻出語となっていた．そのためノード区間が過剰にオーバーラップしてしまい，区間推定が安定しなかった．大語彙であれば，この問題が緩和される事が期待できる．また今回の実装では構造的表象や SSM の比較ではユークリッド距離を用いたが，Muscariello らはヒストグラムを用いた SSM の比較も提案しており [41]，こちらの場合の結果との比較や構造的表象への適用を検証する必要がある．

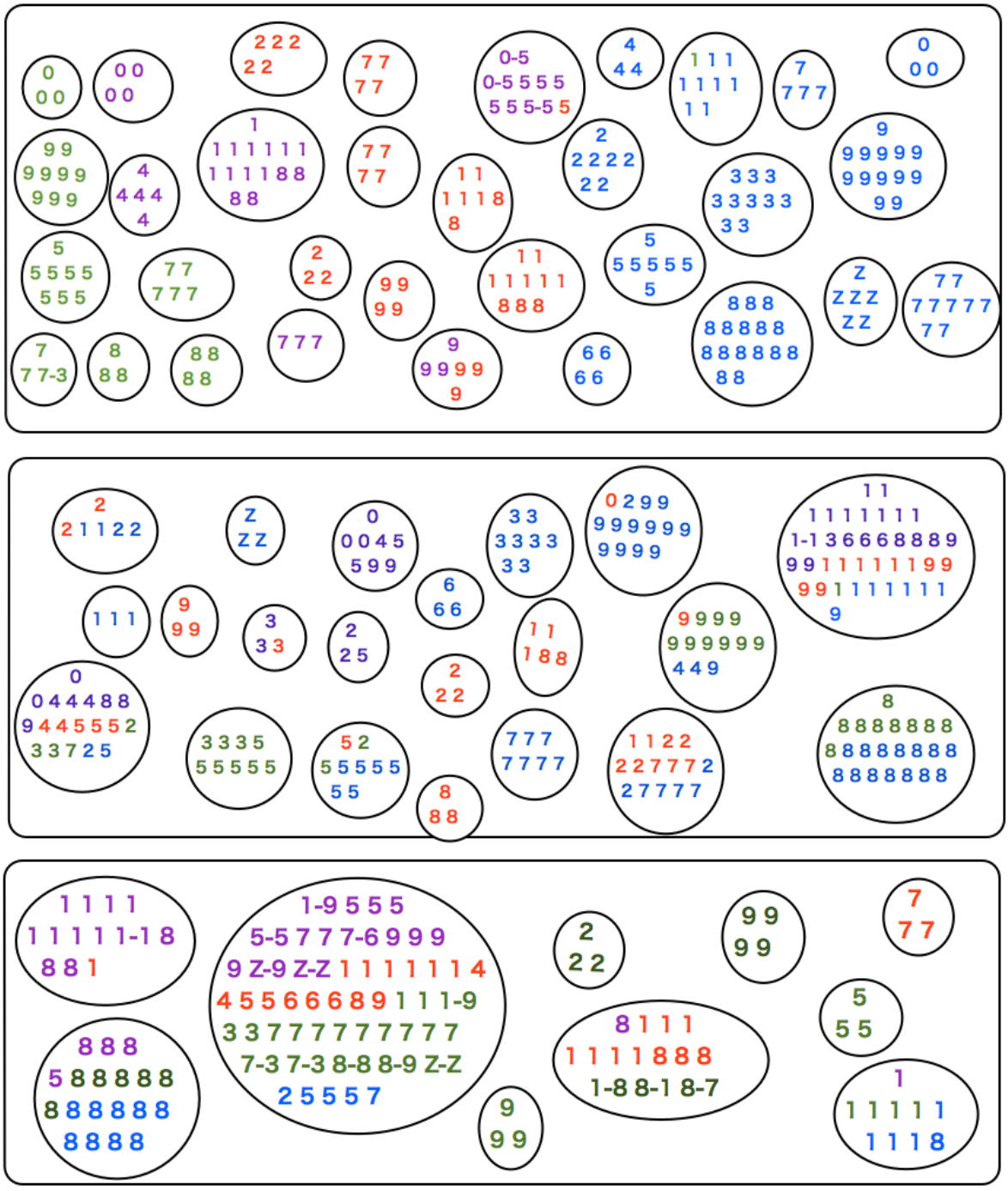

図 5.5: 上から従来手法 1（DTW ベースの距離尺度），従来手法 2（SSM ベースの距離尺度），提案手法（構造的表象ベースの距離尺度）のクラスタリング結果．それぞれ円で囲われた領域を一つのクラスタとする．連続数字音声なのでクラスタ中に表れるノード（単語）は数字 (0 だけは 2 種類の表記 (/rei/, /maru/) があるので後者を “Z” とした) で表し，発話者を色で表す（紫：FAC, 赤：FNG，青：MBD，緑：MAL）．

第6章

結論

6.1 本論文のまとめ

本研究では発達心理学に基づき，言語獲得プロセスを模擬するシステムの特徴を以下のように定義した．

- 連続音声中から言語や音韻などの事前知識を用いずにボトムアップに語彙を獲得する．
- 言語リズム[1]に即した音声処理を行っている．
- 単語全体の語形である語ゲシュタルトを捉えており，非言語情報の変化に頑健である．

従来は，それぞれの観点において (1) 連続音声ストリームからの教師無しパターン抽出，(2) 音声波形からの言語リズム（音節情報）抽出，(3) 非言語情報に対して普遍的構造である音声の構造的表象，のように独立した先行研究（領域）が存在していたが，研究領域において主流であった (1) において (2) や (3) の特徴が考慮される事が少なかった．本研究ではそれらを統合して上述した特徴を兼ね備える音声処理の実現を試み，それによって各々の技術単体で持っていた問題のいくつかに対処できる事を実験で確認した．

実験ではまず，音声波形からの言語リズム（音節核情報）の自動抽出方法について評価実験を行い，主観評価で 88.9%と非常に高い精度（F 値）が得られた．そうして得られた音節核情報を用いて，言語リズムに即した構造的モデリング（音節単位での位置合わせ，モデル化）を提案し，単語認識実験の認識率向上によってその効果を確認した．

最後に連続音声からの教師無し語彙獲得タスクにおいて音声の構造的表象を用いた拡張を行い，複数話者音声に対して頑健に作動する事を確認したが，他の音声正規化手法 (SSM) に対する明確な有意差は得られなかった．

6.2 今後の展望

連続音声からの教師無し語彙獲得タスクにおいて，提案手法はベースラインは大きく上回ったものの，従来手法である SSM を有意に上回る事が出来なかった．理由として考えられるのは，音声の構造的表象を用いて音声照合を行う際に事象間のアライメントが必要となり，これを揃える為に DTW を行った事である．今回のフレームワークでは比較したい音声セグメント間の時間構造のずれが大きく，前処理として DTW での位置合わせが不可欠であったが，その DTW の精度

[1] 本研究では音節の出現間隔でこの言語リズムを定義している

がボトルネックとなっており，そのため同様に前処理に DTW を用いている SSM と似通った結果になってしまったと考えられる．また単語候補のセグメンテーションを行う際に，得られるセグメントの長さをハイパーパラメータで調節する必要があるなどの問題も存在した．一つの解決案としては，これらのフレームワークに言語リズムの考え方を導入する事である．本研究では言語リズムを音節の出現間隔と仮定し，これを自動抽出する手法について検討を行っていたが，これを位置合わせやセグメンテーションに用いる事で，上記の問題を解決できると考えられる．また，今回は連続数字読み上げ音声を用いたが，大語彙での実験も検討課題である．

謝辞

まず本研究を進めるにあたって二年間指導教員として多大なご指導をして頂いた峯松信明教授と広瀬啓吉教授に深く感謝致します．特に峯松信明教授には，常日頃から忙しい中で時間を割いていただき，とても沢山の相談に乗っていただきました．また，日頃の研究活動を支えて下さった高橋登技術専門員，池上恵事務補佐員，折茂結実子事務補佐員にも感謝致します．

また，研究を進めるに当たって日夜議論を交わし，私を導いてくれた元博士課程の鈴木雅之氏[2]と齋藤大輔助教にも感謝します．両者ともに最先端の音声理論から数学の知識，そして広瀬・峯松研究室が長らく取り組んで来た音声の構造的表象など幅広い知識と教養を有し，未熟者であった私に研究の事をはじめ様々な事を教えて頂きました．特に第 4 章の内容は鈴木氏の助力無しでは実現しなかったですし，第 5 章でも齋藤氏のアドバイスが異なったサイズの分布系列間の構造距離を比較する際の大きなヒントとなりました．

苦楽を共にし，時には研究について議論し，時には遊びにと多くの時間を過ごした研究室の方々に感謝します．特に卒論からの 3 年間，様々な事で面倒を見て頂いた研究室の先輩である柏木陽佑氏，橋本浩弥氏，そして苦楽を共にした同期である池島純氏，グエン・ドゥック・ズイ氏，寺井真氏，正木大介氏，毛利圭佑氏には深く感謝しております．彼らの存在無しでは，私がここまで楽しく有意義な学生生活を送れなかったでしょう．そしてこれまでに私を支えていただいた家族，親戚，友人，その他すべての方々にも深く感謝致します．本当にありがとうございました．

2014 年 2 月 6 日
尾崎 洋輔

[2]現 IBM Research.

参考文献

[1] 河原達也, 李晃伸, 伊藤克亘, 小林哲則, 伊藤彰則, 宇津呂武仁, 清水徹, 田本真詞, 荒井和博, 峯松信明, 山本幹雄, 竹沢寿幸, 武田一哉, 松岡達雄, 鹿野清宏, “大語彙日本語連続音声認識研究基盤の整備：評価用連続音声認識プログラムの開発”, 情報通信学会音声言語情報処理研究会，SLP18-1，pp. 1–6, 1997.

[2] N. Minematsu, S. Asakawa, Y. Qiao, D. Saito, and T. Nishimura, “Implementation of robust speech recognition by simulating infants’ speech perception based on the invariant sound shape embedded in utterances,” Proc. Speech and Computer, pp. 35–40, 2009.

[3] O. J. Räsänen, “Fully Unsupervised Word Learning from Continuous Speech Using Transitional Probabilities of Atomic Acoustic Events,” Interspeech’10, Chiba, Japan, pp. 2922–2925, 2010.

[4] J. F. Werker, and H. H. Yeung, “Infant speech perception bootstraps word learning,” TRENDS in Cognitive Sciences, Vol.9, No.11, pp. 519–528, 2005.

[5] I. A. Clemente, M. Heckmann, G. Sagerer, and F. Joublin, “MULTIPLE SEQUENCE ALIGNMENT BASED BOOTSTRAPPING FOR IMPROVED INCREMENTALWORD LEARNING,” ICASSP, pp. 5246–5249 , 2010.

[6] 古井貞熙, “ディジタル信号処理”, 東海大学出版会, 1985.

[7] P. Kuhl, B. Conboy, S. Corina, D. Padden, M. Gaxiola, and T. Nelson, “Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e),” Philosophical transactions of the Royal Society of London , Vol. B, No. 363, pp. 979–1001, 2008.

[8] J. Werker, and S. Curtin, “PRIMIR: A Developmental Framework of Infant Speech Processing,” LANGUAGE LEARNING AND DEVELOPMENT, Vol. 1(2), pp.197–234, 2005.

[9] 加藤正子, “特集にあたって”, コミュニケーション障害学, Vol. 20, No. 2, pp. 84–85, 2003.

[10] 早川勝廣, “ 言語獲得と育児語 ”, 月刊言語, Vol.35, No.9, pp.62–67,2006.

[11] N. トルベツコイ, “音韻論の原理”, 岩波書店, 1958.

[12] P. Eimas, E. Siqueland, P. Jusczyk, and J. Vigorito, “Speech perception in infants,” Science, Vol. 171, pp. 303–306, 1971.

[13] S. Trehub, “The discrimination of foreign speech contrasts by infants and adults,” Child Development, Vol. 47, pp. 466–472, 1976.

[14] 内田伸子，“発達心理学キーワード”，有斐閣双書，2006.

[15] 天野清，“子どものかな文字の習得過程”，秋山書店，1986.

[16] 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，Vol. 20(2), pp. 98–102, 2003.

[17] 高橋登, “幼児のことば遊びの発達：”しりとり”を可能にする条件の分析”，発達心理学研究, Vol. 8(1), pp. 42–52，1997.

[18] 窪薗晴夫，“音声学・音韻論”，くろしお出版，1998.

[19] A. Weibel, “Prosody and speech perception,” PhD Thesis, Carnegie-Mellon University, 1986.

[20] 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.

[21] E. K. Johnson, and P. W. Jusczyk, “Word segmentation by 8-month-olds: When speech cues count more than statistics,” Journal of Memory and Language, Vol. 44, 548–567, 2001.

[22] E. Thiessen, J. R. Saffran, “Spectral tilt as a cue to word segmentation in infancy and adulthood,” Perception and Psychophysics, Vol. 65, pp. 779–791, 2004.

[23] R. Villing, T. Ward, and J. Timoney, “Performance Limits for Envelope based Automatic Syllable Segmentation,” IEE Irish Signals and Systems Conference, pp. 521–525, 2006.

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

[25] R. Villing, J. Timoney, T. Ward, and J. Costello, “Automatic Blind Syllable Segmentation for Continuous Speech,” Irish Signals and Systems Conference Belfast, 2004.

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

[27] A. Galves, J. Garcia, D. Duarte, and C. Galves, “Sonority as a basis for rhythmic class discrimination,” Speech Prosody, Aix-en-Provence, 2002.

[28] W. M. NG, “Spoken Language Identification withProsodic Features,” 香港中文大学博士論文, 2011.

[29] 中村健太郎，“音のしくみ”，ナツメ社，1999.

[30] A. Bell, J. H. Greenberg, “Syllabic consonants,” Universals of Human Language, 1978.

[31] 斎藤純男，“日本語音声学入門”，三省堂，1997.

[32] 窪薗晴夫，本間猛，“音節とモーラ”，研究社，2002.

[33] 峯松信明，“音声の音響的普遍構造の歪みに着眼した外国語発音の自動評定”，電子情報通信学会音声研究会，Vol. 180，pp.31–36，2003.

[34] A. Cros, D. Demolin, A. Flesia, and A. Galves, “On the relationship between intra-oral pressure and speech sonority,” INTERSPEECH, pp. 2165–2168, 2005.

[35] M. Pitz, and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space,” IEEE Trans.Speech and Audio Processing, Vol. 13, pp. 930–944, 2005.

[36] 高澤真章，峯松信明，広瀬啓吉，“CALL 応用を目的とした教師・学習者の発声間における時間アライメントに関する実験的検討”，日本音響学会春季講演論文集，1-P-12，pp. 429–432，2010.

[37] 村上隆夫，峯松信明，広瀬啓吉，“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，Vol. J91-A，No. 2，pp. 181–191，2008.

[38] S. Asakawa N. Minematsu, and K. Hirose, “Multi-stream parameterization for structural speech recognition,” ICASSP, pp. 4097–4100, 2008.

[39] A. S. Park, and J.Glass, “Unsupervised pattern discovery in speech,” IEEE transactions on audio speech and language processing, 16, 1, pp.186–197, 2008.

[40] F. McInnes, and S. Goldwater, “Unsupervised extraction of recurring words from infant-directed speech,” Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Boston, MA, pp. 2006–2012, 2011.

[41] A. Muscariello , G. Gravier, and F. Bimbot, “Towards robust word discovery by self-similarity matrix comparison”. ICASSP, pp. 5640–5643, 2011.

[42] Y. Zhang, and J. Glass, “Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams,” ASRU, pp. 398–403, 2009.

[43] Y. Wang, and L. Lee, “Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning,” ICASSP, pp. 8232–8236, 2013.

[44] M. Newman, “Fast algorithm for detecting community structure in networks.” Physical review, pp. 69–74, 2004.

発表文献

国際会議論文

[1] N. Minematsu, Y. Ozaki, K. Hirose, D. Erickson, “Speaker-invariant and rhythm-sensitive representation of spoken words,” Proc. APSIPA (CD-ROM), 2013.

[2] J. Williams, D. Erickson, Y. Ozaki, A. Suemitsu, N. Minematsu, and O. Fujimura, “Neutralizing differences in mandible displacement for English vowels”, In Proceedings of Meetings on Acoustics, Vol. 19, 2013.

国内研究会・全国大会

[3] 尾崎洋輔，峯松信明，広瀬啓吉，Donna Erickson，“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”，日本音響学会秋季講演論文集，2-Q-a4，pp. 519–522，2012.

[4] 尾崎洋輔，峯松信明，広瀬啓吉，Donna Erickson，“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の初期検討”，電子情報通信学会音声研究会資料，SP2012-94，pp. 113–118, 2012.

[5] J. Williams, D. Erickson, Y. Ozaki, A. Suemitsu, N. Minematsu, and O. Fujimura, “Neutralizing differences in jaw displacement for English vowels,” The Journal of the Acoustical Society of America, 133(5), pp. 3607–3607, 2013.

[6] 尾崎洋輔，柏木陽佑，齋藤大輔，峯松信明，広瀬啓吉，“音声雑音に頑健な主話者区間検出に関する検討”，日本音響学会秋季講演論文集，1-P-20d，pp. 151–154，2013.

[7] C. Zhang, Y. Ozaki, D. Saito, N. Minematsu and K. Hirose, “Use of invariant and structural features in discriminative models for speech recognition,” Autumn Meeting of Acoustical Society of Japan, 1-P-26b, pp.163-164, 2013.

[8] 尾崎洋輔，柏木陽佑，齋藤大輔，峯松信明，広瀬啓吉，“音声の構造的表象による頑健な教師無し語彙獲得に関する実験的検討”，日本音響学会春季講演論文集，3-Q5-10，2014（発表予定）.

学位論文

[9] 尾崎 洋輔，“波形包絡を用いた音節核の自動抽出と音声認識への応用”，東京大学工学部電子工学科卒業論文，2012.