HMM 音声合成のための H/L 型アクセント推定を統合した音響モデリング

学籍番号　２３４１３５０２　氏名　足立 貴昭
指導教員名　徳田 恵一

1　はじめに

近年，カーナビやスマートフォンなどで，合成音声が身近になり，システムとコミュニケーションを取る上で，人間のような自然なアクセントを表現する合成音声が求められている．

隠れマルコフモデル (Hidden Markov Model; HMM) に基づく音声合成システム [1] では，統計モデルにより小規模データベースから比較的安定した音声合成が可能である．このシステムでは，テキストからアクセント情報を取得する必要があり，従来は規則に基づく手法が用いられている．しかし，単語毎の辞書構築が困難であり，例外的なアクセント型の推定ができないため，近年は統計的手法が注目されている．一方，従来の音声合成システムの多くは，アクセント推定と音響モデル学習が独立して行われており，テキストから音声波形を生成する問題を直接定式化できていない．

そこで本研究では，統計モデルである条件付確率場 (Conditional Random Fields; CRF) に基づき，新たに推定単位をモーラとした H/L 型アクセント推定を提案する．また，H/L 型アクセント推定と音響モデル学習を統合したモデル構造を定義し，相互の影響を考慮したモデル学習を行うことで，音声合成のアクセントの自然性向上が期待できる．

2　アクセント推定

日本語単語のアクセントは，モーラと呼ばれる仮名 1 文字に対応する単位におけるピッチの高低で表現でき，その上昇や下降パターンにより，図 1 に示すようなアクセント型に分類される．各単語は固有にアクセント型を持つが，文章中においてアクセント型が変化することがある．そのため，従来の音声合成システムでは，匂坂らによるアクセント結合規則 [2] により，文中でのアクセント型変化を推定している．しかし，規則に基づく手法は，予め単語辞書に規則を記述する作業が困難であり，例外的なアクセント型変化に対応できないと言う問題がある．そこで，近年は CRF などの統計モデルによるアクセント推定が注目されている．

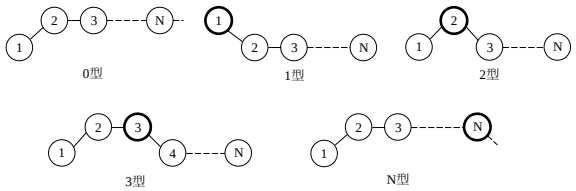

図 1: アクセント型の例

3　条件付確率場に基づく H/L 型アクセント推定

条件付確率場 (CRF) に基づく H/L 型アクセント推定では，モーラ単位においてアクセントの高低 (High/Low) を推定する．CRF に基づき，モーラ系列 $\boldsymbol{W} = \{\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_M\}$ が与えられた時に，各モーラの H/L 型アクセントラベル系列 $\boldsymbol{L} = \{l_1, l_2, \ldots, l_M\}$ を出力する確率を式 (1) により定義する．

$$P(\boldsymbol{L}|\boldsymbol{W}, \boldsymbol{\lambda}_L) = \frac{1}{Z(\boldsymbol{W})} \exp\left\{\sum_{m=1}^{M}\sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{m}_m, l_m)\right\} \quad (1)$$

ただし，$Z(\boldsymbol{W})$ は正規化項，f_k は素性関数，λ_k は素性関数 f_k の重み，K は素性関数の数，M は系列の長さである．素性関数 f_k は，ある特徴を満たすときに 1 を，それ以外のときに 0 を返す関数であり，全データ中にその特徴が出現する頻度を表現する．ここでは，アクセント変化を表現する特徴として，40 種類の形態素解析情報を組み合わせることで素性関数を定義し，アクセント推定に有効な特徴の調査を行う．

CRF の学習は，学習データに対する対数尤度 $\log P(\boldsymbol{L}|\boldsymbol{W}, \boldsymbol{\lambda}_L)$ が最大となるようなモデルパラメータ $\boldsymbol{\lambda}_L$ を，最尤推定に基づいて選択することで行う．ただし，モデルパラメータ $\boldsymbol{\lambda}_L$ を解析的に求めることは困難であるため，準ニュートン法によるパラメータ推定を行う．H/L 型アクセント推定では，学習で得られたモデルにおいて，モーラ系列 $\boldsymbol{W}$ に対して出力確率が最大となる H/L アクセントラベル系列 $\boldsymbol{L}$ を求める．

4　H/L 型アクセント推定を統合した音響モデリング

H/L 型アクセント推定と音響モデルの学習を相互に考慮するため，両モデルを統合した新しいモデル構造を定義する．

観測ベクトル $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ が与えられた時，音響モデル $\boldsymbol{\lambda}_H$ の尤度関数は式 (2) で与えられる．ただし，隠れ変数 $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ は状態系列である．

$$P(\boldsymbol{O}|\boldsymbol{L}, \boldsymbol{\lambda}_H) = \sum_{\boldsymbol{q}} P(\boldsymbol{O}|\boldsymbol{q}, \boldsymbol{\lambda}_H) P(\boldsymbol{q}|\boldsymbol{L}, \boldsymbol{\lambda}_H) \quad (2)$$

一方，アクセント推定モデル $\boldsymbol{\lambda}_L$ の尤度関数 $P(\boldsymbol{L}|\boldsymbol{W}, \boldsymbol{\lambda}_L)$ は式 (1) で与えられている．これらの音響モデルとアクセント推定モデルを 1 つのモデルとみなすと，モーラ系列 $\boldsymbol{W}$ が与えられたときに観測ベクトル $\boldsymbol{O}$ を直接出力する統合モデル $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_H, \boldsymbol{\lambda}_L\}$ の尤度関数は式 (3) で表される．

$$P(\boldsymbol{O}|\boldsymbol{W}, \boldsymbol{\lambda}) = \sum_{\boldsymbol{L}}\sum_{\boldsymbol{q}} P(\boldsymbol{O}|\boldsymbol{q}, \boldsymbol{\lambda}_H) P(\boldsymbol{q}|\boldsymbol{L}, \boldsymbol{\lambda}_H) P(\boldsymbol{L}|\boldsymbol{W}, \boldsymbol{\lambda}_L) \quad (3)$$

統合モデルの学習は，EM アルゴリズム [3] により行い，$\mathcal{Q}$ 関数を用いて尤度関数 $P(\boldsymbol{O}|\boldsymbol{W}, \boldsymbol{\lambda})$ が最大となるようにモデルパラメータ $\boldsymbol{\lambda}$ の更新を行う．統合モデルの $\mathcal{Q}$ 関数は式 (4) で表される．ただし，$\boldsymbol{\lambda}'$ は更新後のモデルパラメータである．

$$\mathcal{Q}(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \sum_{\boldsymbol{L}}\sum_{\boldsymbol{q}} P(\boldsymbol{q}, \boldsymbol{L}|\boldsymbol{O}, \boldsymbol{W}, \boldsymbol{\lambda}) \log\left[P(\boldsymbol{O}|\boldsymbol{q}, \boldsymbol{\lambda}'_H) P(\boldsymbol{q}|\boldsymbol{L}, \boldsymbol{\lambda}'_H) P(\boldsymbol{L}|\boldsymbol{W}, \boldsymbol{\lambda}'_L)\right] \quad (4)$$

EM アルゴリズムでは，E ステップにおいて $\mathcal{Q}$ 関数の計算，M ステップにおいて統合モデルのパラメータ更新を行い，この 2 ステップを繰り返し行う．ただし，M ステップでは，音響モデル $\boldsymbol{\lambda}_H$ とアクセント推定モデル $\boldsymbol{\lambda}_L$ を個別に更新する．

4.1　音響モデルの更新

音響モデル $\boldsymbol{\lambda}_H$ の更新は，HMM の各モデルパラメータについて $\mathcal{Q}$ 関数を微分することで最適なパラメータを求める．例えば，出力確率分布（ガウス分布）の平均ベクトル $\boldsymbol{\mu}_i$ と共分散行列 $\boldsymbol{\Sigma}_i$ の更新式は次式で与えられる．

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_t \sum_{\boldsymbol{L}} \gamma_i(t, \boldsymbol{L}) \boldsymbol{o}_t}{\sum_t \sum_{\boldsymbol{L}} \gamma_i(t, \boldsymbol{L})} \quad (5)$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{\sum_t \sum_{\boldsymbol{L}} \gamma_i(t, \boldsymbol{L}) (\boldsymbol{o}_t - \boldsymbol{\mu}_i)(\boldsymbol{o}_t - \boldsymbol{\mu}_i)^{\mathrm{T}}}{\sum_t \sum_{\boldsymbol{L}} \gamma_i(t, \boldsymbol{L})} \quad (6)$$

ただし，$\gamma_i(t, \boldsymbol{L})$ は次式として定義する．

$$\begin{aligned}\gamma_i(t, \boldsymbol{L}) &= P(q_t = S_i, \boldsymbol{L}|\boldsymbol{O}, \boldsymbol{W}, \boldsymbol{\lambda}) \\ &= P(q_t = S_i|\boldsymbol{L}, \boldsymbol{O}, \boldsymbol{\lambda}_H) P(\boldsymbol{L}|\boldsymbol{O}, \boldsymbol{W}, \boldsymbol{\lambda}) \end{aligned} \quad (7)$$

ここで，$P(q_t = S_i|\boldsymbol{L}, \boldsymbol{O}, \boldsymbol{\lambda}_H)$ は，HMM にラベル系列 $\boldsymbol{L}$ を与えたときの，状態 q_t の事後確率であり，Forward-Backward アルゴリズムにより計算できる．また，$P(\boldsymbol{L}|\boldsymbol{O}, \boldsymbol{W}, \boldsymbol{\lambda})$ はラベル系列 $\boldsymbol{L}$ の事後確率であり，式 (8) により計算される．

$$P(\boldsymbol{L}|\boldsymbol{O}, \boldsymbol{W}, \boldsymbol{\lambda}) = \frac{P(\boldsymbol{L}|\boldsymbol{W}, \boldsymbol{\lambda}_L) P(\boldsymbol{O}|\boldsymbol{L}, \boldsymbol{\lambda}_H)}{\sum_{\boldsymbol{L}'} P(\boldsymbol{L}'|\boldsymbol{W}, \boldsymbol{\lambda}_L) P(\boldsymbol{O}|\boldsymbol{L}', \boldsymbol{\lambda}_H)} \quad (8)$$

4.2 H/L 型アクセント推定モデルの更新

H/L 型アクセント推定モデル $\boldsymbol{\lambda}_L$ の更新は，$\mathcal{Q}$ 関数における $\boldsymbol{\lambda}_L$ に関する項について，CRF の学習を $P(\boldsymbol{L}|\boldsymbol{O}, \boldsymbol{W}, \boldsymbol{\lambda})$ で重み付けすることで行う．

$$\hat{\boldsymbol{\lambda}}_L = \arg\max_{\boldsymbol{\lambda}_L} \sum_{\boldsymbol{L}} P(\boldsymbol{L}|\boldsymbol{O}, \boldsymbol{W}, \boldsymbol{\lambda}) \log P(\boldsymbol{L}|\boldsymbol{W}, \boldsymbol{\lambda}'_L) \quad (9)$$

4.3 統合モデルの EM アルゴリズム

統合モデルの EM アルゴリズムでは，全てのラベル系列 $\boldsymbol{L}$ について，事後確率 $P(\boldsymbol{L}|\boldsymbol{O}, \boldsymbol{W}, \boldsymbol{\lambda})$ を計算する必要がある．しかし，実際には計算量の観点から実現が困難である．そこで，本研究では N-best 近似を用い，ラベル系列 $\boldsymbol{L}$ を計算可能な数に制限して学習を行う．すなわち，EM アルゴリズムの E ステップにおいて，まず，アクセント推定モデル $\boldsymbol{\lambda}_L$ に従ってラベル系列 $\boldsymbol{L}_n, n = 1, \ldots, N$ を出力し，N 個のラベル系列について事後確率 $P(\boldsymbol{L}_n|\boldsymbol{O}, \boldsymbol{W}, \boldsymbol{\lambda})$ を計算することで，$\mathcal{Q}$ 関数を求める．また，M ステップにおいても N 個のラベル系列を用いて，各モデルパラメータの更新を行う．

このように，統合モデルでは音響モデルと H/L 型アクセント推定モデルを相互に考慮し，交互にモデルパラメータを更新する．そのため，それぞれのモデルが最適化され，音声合成のアクセントの自然性が向上することが期待できる．

5 評価実験

5.1 アクセント推定実験

CRF に基づく H/L 型アクセント推定を評価するため，アクセント推定実験を行った．CRF の学習には音素バランス文 426 文（13803 モーラ），各手法の評価には 53 文（1223 モーラ）を用いた．素性関数は，まず形態素解析情報を 1 種類から 7 種類まで組み合わせた特徴を定義し，各組み合わせ毎に，正解率を調査した．次に各組み合わせ数 X において，上位複数個の素性関数を同時に使用したアクセント推定を行い，それぞれにおいて最も正解率が高い結果を比較対象（CRF_X）とし，従来法である規則に基づくアクセント処理（$RULE$）との比較を行った．尚，単語単独でのアクセントを推定前アクセント（PRE）として算出した．アクセント推定の正解率を図 2 に示す．図 2 を見ると，$RULE$ に比べて各 CRF_X の正解率が高い傾向にあり，提案法によるアクセント推定の有効性が確認できた．ただし，CRF_7 では正解率が低下しており，これは素性関数の複雑化により，特徴を満たす学習データの減少が原因として考えられる．そのため，学習データを増やすことで更なる推定精度の向上が期待できる．

5.2 音声合成実験

統合モデルによる音声合成を評価するため，MOS に基づく主観評価実験を行った．実験データは男性話者 1 名による音声データ 479 文，サンプリング周波数 48kHz を用い，この内，学習データに 426 文，評価データに 53 文を用いた．特徴量としては，スペクトルパラメータを 39 次元の STRAIGHT メルケプストラム，基本周波数パラメータを対数基本周波数とし，それぞれの静的特徴量と Δ, Δ^2 を用いた．尚，HMM は 5 状態スキップなし left-to-right 型 HMM とした．

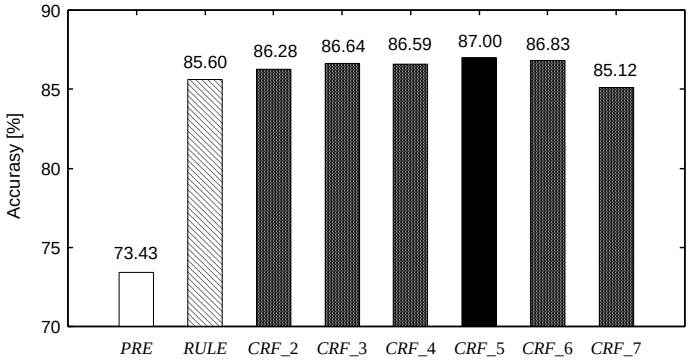

図 2: アクセント推定の正解率

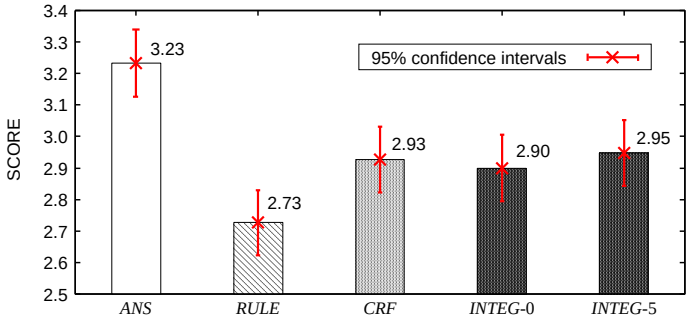

図 3: 主観評価実験の結果

統合モデルは，モデルの繰り返し学習の回数が異なる 2 手法とし，繰り返し学習を行わない ($INTEG$-0)，繰り返し学習を 5 回行う ($INTEG$-5) である．従来法である独立モデルは，アクセント推定手法が異なる 3 手法とし，正解アクセントを与える (ANS)，規則に基づくアクセント処理を行う ($RULE$)，CRF に基づく H/L 型アクセント推定を行う (CRF) である．これら 5 手法について，被験者 10 名に対し主観評価実験を行った．各被験者は，各手法において評価文 53 文からランダムに 15 文を受聴し，合成音声のアクセントの自然性を 1 点から 5 点で評価した．主観評価実験の結果を図 3 に示す．図 3 を見ると，$RULE$ に比べて CRF が高い評価となり，アクセント推定の精度向上によって音声の自然性が向上することが確認された．また，$INTEG$-5 が CRF より高い評価を得ており，これは統合モデルを繰り返し学習する提案法によって，両モデルを相互に考慮した最適化が行われ，音声合成のアクセントの自然性が向上したと考えられる．

6 むすび

本研究では，CRF に基づく H/L 型アクセント推定，及び H/L 型アクセント推定を統合した音響モデリングを提案した．実験結果により，提案法による音声合成のアクセントの自然性向上を確認した．今後，学習データを増加し，統合モデルの学習を繰り返すことで，更なる自然性向上が期待される．

参考文献

[1] 吉村貴克, 徳田恵一, 益子貴史, 小林隆夫, 北村正, “HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化”, 信学論, vol. J83–D–II, no. 11, pp. 2099–2107, Nov. 2000.

[2] 匂坂芳典, 佐藤大和, “日本語単語連鎖のアクセント規則”, 電子通信学会論文誌, Vol.J66-D, No.7, pp.849-856, July 1983

[3] A.P.Dempster, et al., “Maximum-likelihood from incomplete data via the EM algorithm”, J.Royal Statist. Soc. Ser.B(methodological), vol. 39, pp. 1–38, 1977.