条件付確率場に基づく基本周波数推定

学籍番号　２３４１３５０３　氏名　天野　貴裕
指導教員名　南角　吉彦

1　はじめに

音高（ピッチ）を表現する基本周波数 (F_0) は，音声情報処理技術に広く用いられる重要な音響特徴量である [1]．したがって高い抽出精度が求められており，様々な特徴量を用いて F_0 推定が行われている．しかし，環境に合わせた調節が人手によって必要な手法が多く，未だ確立された手法は提案されていない．そこで条件付確率場 (Conditional Random Fields; CRF) [2] に基づく F_0 推定を提案する．CRF は自由な構造を表現可能な統計モデルであり，複数の特徴量を選択的に利用するため高い識別能力を持つ．これにより高精度な F_0 推定を目指す．

2　基本周波数

音声は F_0 の有無により，母音などを発する際の音である有声音と，破裂音や摩擦音，ひそひそ声を発する際の音である無声音に大別される．また F_0 は人の聴覚において音のピッチを司っており，語句のアクセントやイントネーション，男女間や年齢における声の高さの違い，感情による音高の変化を表現し，音声の大量の情報を含んだ重要な特徴である．

F_0 の推定には様々な特徴量が利用可能である．しかし，音声波形は準周期信号であることや，雑音による影響が大きいなどの理由によって，F_0 抽出時に有声・無声誤りや倍ピッチ誤りが発生してしまう．そのため従来の推定法では，複数の特徴量の利用，雑音環境に合わせた調節を人手で行うことで頑健性を向上させており，どのような環境下においても高い推定精度をもった手法は未だ確立されていない．

3　条件付確率場

CRF では観測系列 $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$ が入力された際に，ラベル列 $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$ を出力する条件付確率を式 (1) で表現する．

$$P(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\Lambda}) = \frac{1}{Z} \exp \sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{x}, \boldsymbol{y}) \tag{1}$$

$$Z = \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \exp \sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{x}, \boldsymbol{y}) \tag{2}$$

ただし， $f_k(\cdot)$ は素性番号 k によって与えられる素性関数であり， λ_k は素性関数 $f_k(\cdot)$ にかかる重みパラメータ， K は素性の数である．また，Z は $\mathcal{Y}(\boldsymbol{x})$ は全種類のラベルの組み合わせを考慮した正規化項を表している．素性関数は入力系列と出力系列のある関係性を表現したものであり，式 (3) のようになる．

$$f(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 1 & : (\boldsymbol{x}, \boldsymbol{y}) \text{ がある条件を満たすとき} \\ 0 & : \text{otherwise} \end{cases} \tag{3}$$

CRF は関係性の重要度である重みパラメータに対して事後確率を最大化するように学習が行われるため，識別に有効な情報を選択的に利用し高い認識性能を実現している．

4　条件付確率場に基づく基本周波数推定

F_0 推定を CRF によりモデル化する．モデル化によって F_0 についての情報がデータから自動学習され，環境に合わせた人手による調節が不要となる．さらに複数特徴量を識別的に扱うことが可能となり，推定精度の向上が期待できる．

モデル化にあたって，提案法の入出力系列について述べる．F_0 推定のための特徴量は連続値であるため，入力系列 $\boldsymbol{X}$ を D 次元の特徴量ベクトル系列として以下のように定義する．

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T) \tag{4}$$

$$\boldsymbol{x}_t = [x_{t1}\ x_{t2}\ \cdots\ x_{tD}]^\top \tag{5}$$

$\boldsymbol{x}$ はどのような特徴量でも与えることが可能であり，数種類の特徴量を並べて定義することも可能である．また CRF において出力系列は離散値であるが，F_0 系列は連続値であるため，F_0 系列を離散化する．サンプリング周波数 f_s で離散化された音声 s_n の時刻 t における F_0 は，基本周期を示すサンプル数 L によって， $F_0 = f_s/L$ と計算されるため，F_0 系列を離散な基本周期の系列として離散化した．このとき基本周期を持たない無声音のラベルも設定することで，有声・無声判別も同時に行っている．これらの入出力系列の条件付確率を，F_0 の時系列による変化を考慮するための遷移素性 (a) と時刻における入出力の関係を考慮するための出力素性 (b) を用いて式 (6) で定義する．

$$P(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\Lambda}) = \frac{1}{Z} \exp \left[\sum_{t=1}^{T} \left\{ \sum_{k=1}^{K^{(a)}} \lambda_k^{(a)} f_k^{(a)}(\boldsymbol{x}_t, y_{t-1}, y_t) + \sum_{k=1}^{K^{(b)}} \lambda_k^{(b)} f_k^{(b)}(\boldsymbol{x}_t, y_t) \right\} \right] \tag{6}$$

ただし， T は全データの総フレーム数， $(\cdot)^{(a)}$ は遷移素性，$(\cdot)^{(b)}$ は出力素性に関わるものである．

素性関数は，$f_k^{(a)}$ の「ある時刻において k によって決まる出力ラベルの組 $\langle (y, y') \rangle_k$ で遷移が起きたとき， k によって決まる特徴量ベクトルの $\langle d \rangle_k$ 次元の値を返す」という式 (7) のものと，$f_k^{(b)}$ の「ある時刻において k によって決まる入力特徴量ベクトルに対してラベル $\langle y \rangle_k$ が出力されたとき，k によって決まる特徴量ベクトルの $\langle d \rangle_k$ 次元の値を返す」という式 (8) のもの，遷移・出力素性ともに特徴量の値ではなく式 (3) のような 1 を返すものの全 4 種を用いた．

$$f_k^{(a)}(\boldsymbol{x}_t, y_{t-1}, y_t) = \begin{cases} x_{t,\langle d \rangle_k} & : (y_{t-1}, y_t) = \langle (y, y') \rangle_k \\ 0 & : \text{otherwise} \end{cases} \tag{7}$$

$$f_k^{(b)}(\boldsymbol{x}_t, y_t) = \begin{cases} x_{t,\langle d \rangle_k} & : y_t = \langle y \rangle_k \\ 0 & : \text{otherwise} \end{cases} \tag{8}$$

また，F_0 推定手法の一つとして広く用いられている RAPT [3] の枠組みを，F_0 推定のための有用な事前知識として，今回のモデル構造に反映させた．

5　評価実験

提案法の有効性を示すために，F_0 推定の評価実験を行った．学習データは ATR 日本語音声 DB B-set 男性話者 mht から 450 文章を用い，open テストに 450 文以外から 53 文章を，close テストとして学習データから 50 文章を用いた．このとき出力ラベルは DB 中の基本周期の最大と最小のものを利用し，最小値のラベル (37) と最大値のラベル (539) から有声音で 503 ラベル，さらに無声ラベルを追加した全 504 ラベルを識別する．

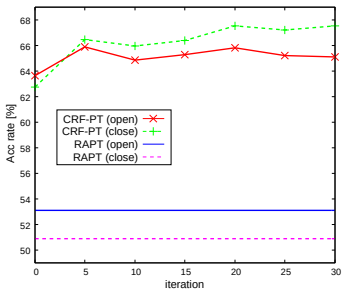


図 1: Acc rate

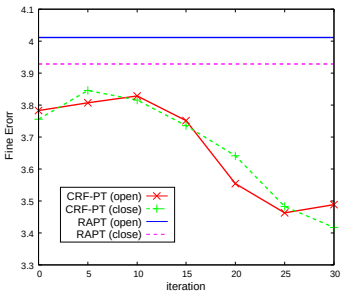


図 2: Fine Error

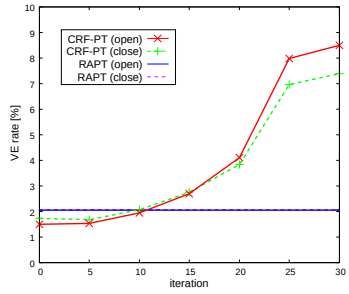


図 3: VE rate

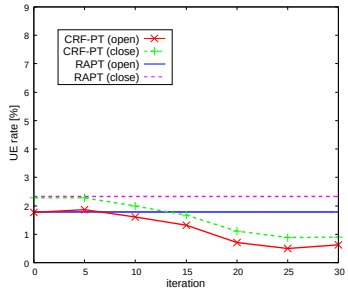


図 4: UE rate

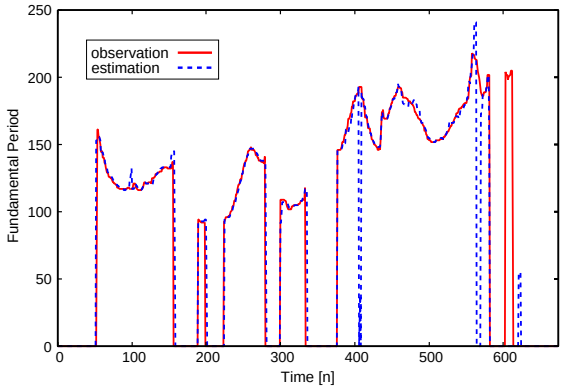
従来法の RAPT と提案法である CRF-PT の学習回数 5 回毎のモデルにおいて F_0 推定を行い，以下の評価値における結果をグラフにまとめた．図 1〜図 4 にて示す．

- Accuracy (Acc) rate
  正解と完全に一致したフレームの割合　（図 1）
- Fine Erorr
  推定誤差 10%以内のフレームの平均二乗偏差　（図 2）
- Voice to Unoice Error (VE) rate
  有声音を無声音に間違えたフレームの割合　（図 3）
- Unvoice to Voice Error (UE) rate
  無声音を有声音に間違えたフレームの割合　（図 4）

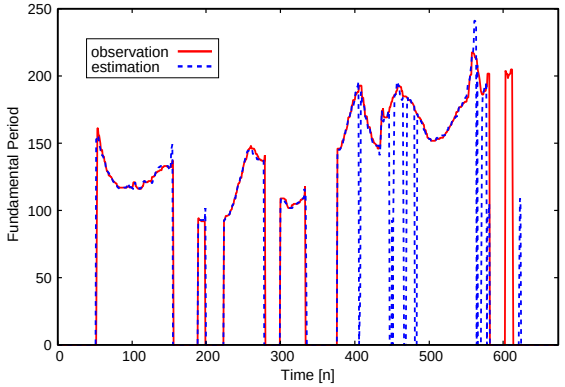
また図 5 に，ある open データに対する推定結果の基本周期を示す．図 5 では赤線が実際に観測される F_0 系列を，青線が推定した F_0 系列を表している．

図 1 において，全学習回数において提案法が従来法より高い性能を示した．しかし，学習回数による改善はあまり見られていない．CRF において close データに対する Acc rate はすぐに 100%近くになるという傾向があることが知られているが今回はその傾向が観測されず，今回用いたモデル構造が識別に有効でなかった可能性が考えられる．つまり，今回与えたモデル構造は問題空間に対し線形な識別境界を引くものであるが，この識別境界の表現能力が F_0 の推定問題に対して不十分であったためこのような結果となったと考えられる．したがって F_0 の識別問題は想定していたものほど単純ではなく，識別境界の表現能力を特徴量を増やすことや非線形なものへと拡張することにより高める必要があるということが分かった．

また，図 2 では提案法の学習が進むにつれて有声音における推定誤差の平均二乗偏差が下がっていくため，学習による性能の向上が確認された．したがって，図 1 の結果となった理由は，有声・無声誤りによる影響が大きいからだといえる．実際に図 3，図 4 を観察すると，学習が進むに伴い有声音に対して無声音が出力されやすくなっていることが分かる．これは，事後確率を直接モデル化している CRF において，学習データ中に観測される有声音のある 1 ラベルより，より多く観測される無声音のラベルの出現確率を大きくするようにして学習が進んでしまったためだと考えられる．以上の考察は，図 5 においても確認することが可能である．今回のモデル構造では有声音の 1 クラスと無声音の 1 クラスを同等のものとして扱っているため，無声音のラベルに対する重みを設定するなどといった対処により改善が期待される．



(a) 学習前の初期モデル



(b) 学習回数 20 回のモデル

図 5: F_0 系列（赤線）と推定結果（青線）

6　むすび

本研究では，CRF に基づく F_0 推定を提案した．モデル化を行うことで，データ中の F_0 についての情報を自動学習し，従来の人手による環境に合わせた調節が不要となる．また，識別モデルである CRF を用いることで推定に有効な情報を選択的に利用し，さらに時系列を考慮した F_0 推定が行える．これによって頑健で高精度な F_0 推定が実現可能である．F_0 の推定実験により提案法が従来法と同等以上の識別性能を示すことを確認した．また有声と無声の判別に課題が残るが，学習による識別性能の向上が確認され，特に有声音の識別において大きな改善がみられた．

今後の課題としては，使用する特徴量の検討，無声音のラベルに対して重みを設定することが挙げられる．

参考文献

[1] 鈴木 久喜，“ピッチ抽出の今昔，” 日本音響学会誌 56 巻 2 号，pp.121–128, 2000.

[2] J.Lafferty, A.McCallum, and F.Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” Proc. of ICML, pp.282–289, 2001.

[3] D.Talkin, “A Robust algorithm for Pitch Tracking,” Speech Coding Synthesis, pp.495–518, 1995.