幼児音声の基本周波数および有声区間の推定法*

○中谷 智広† 天野 成昭† 入野 俊夫†‡
(† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所 音声オープンラボ ‡CREST)

1 はじめに

幼児が言語を学習する過程においてプロソディ情報は重要な役割を果たすと考えられている。この情報の分析には、親子の会話の有声/無声区間 (V/UV) および基本周波数 (F_0) のデータが必要であろう。しかし、幼児音声の音響的特徴に関する従来研究は少なく、大量データに対して、これらの値を推定したという報告はない。本稿では、新たに調波成分の占有度を用いた V/UV 判定法を提案するとともに、幼児音声の F_0 推定に瞬時周波数を用いる方法 [1] を適用する。

2 幼児音声の特徴と課題

親子の会話を継続的に録音した音声データ（5 組、0〜5 歳）[2] を調べると、幼児音声は、成人とは異なる以下の音響的特徴を持つといえる。

1. F_0 のとる範囲の広さ (200 ∼ 2000 Hz)
2. F_0 の不連続性（2 倍もしくは 1/2 倍に突然遷移）
3. 高域まで強いエネルギを持ちうる有声音の特性

また、分析に用いたデータは日常生活の録音であるため、録音条件が必ずしも一定しておらず、一般的に SNR が低い（約 20〜-5dB）。

このような特徴は、音声の F_0 および V/UV 推定性能を劣化させる要因となる。F_0 探索範囲の拡大は、いわゆる「半ピッチ・倍ピッチエラー」につながり、F_0 の不連続性は、平滑化処理だけではこのエラーに対処できないことを意味する。また、成人の有声音は、低域の信号パワーに比べて高域が小さいことが一つの判定基準である [3] が、幼児音声ではこれは成り立たない。さらに、従来の V/UV 判定法は、必ずしも信頼性の高くない複数の判定基準（ケプストラムのピーク値、パワー、zero crossing 比など）を発見的に組み合わせることで精度を向上させていた [4] が、このような方法は録音条件 (信号のレベルなど) の変化に対して頑健ではない。また、そもそも低 SNR 条件下では、個々の特徴量を正確に求めるのは難しくなる。

3 瞬時周波数を用いた推定法

低 SNR 条件下でも頑健かつ正確に調波成分の特徴量を抽出するために、我々は占有度に基づく分析手法を提案している [1]。本稿でも、この手法を用いて推定性能の向上をはかる。占有度による方法では、雑音下で占有的なパワーを持つ調波成分を特定し、その特徴量のみを用いて特徴抽出を行う。すでに、成人音声については、占有度を用いて高性能な F_0 推定法が構成できることが示されている。

*Fundamental frequency and voiced segment estimation in infant utterance based on dominant harmonic components, by Nakatani, T., Amano, S. (NTT Corporation), and Irino, T. (NTT Corporation/CREST)

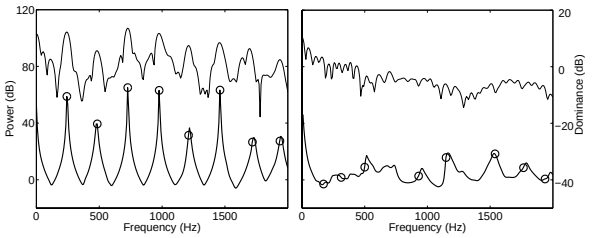

図 1: 幼児音声の有声区間 (左) と無声区間 (右) の占有度 (太線) と対数パワースペクトル (細線) と不動点 (円)

3.1 占有度

瞬時周波数 $\dot{\phi}(\omega)$ は、雑音の影響の少ない調波成分の周波数近傍で一定値となる性質を持つ。この一定度合いを評価する尺度が占有度 $D_0(\omega_c)$ であり、各周波数 bin (中心周波数 ω_c) ごとに以下で定義される。

$$D_0(\omega_c) = 10\log_{10}(1/B(\omega_c)^2), \quad (1)$$

$$B(\omega_c)^2 = \frac{\int_{\omega_c-\triangle\omega/2}^{\omega_c+\triangle\omega/2}(\dot{\phi}(\omega)-\omega_c)^2 S(\omega)^2 \mathrm{d}\omega}{\int_{\omega_c-\triangle\omega/2}^{\omega_c+\triangle\omega/2} S(\omega)^2 \mathrm{d}\omega}. \quad (2)$$

占有度は、パワースペクトル $S(\omega)^2$ で正規化した値でもあるため、非調波成分に対する調波成分の相対的な強度のみを評価するものである。信号の絶対的なパワーに依存せず、ほぼ同じ範囲の値をとる (約 $-40 \sim 0$dB)。

図 1 に、幼児音声の有声/無声区間の占有度を各周波数 bin ごとにプロットした結果を示す。図より、有声区間では、占有度は各調波成分に相当する規則的な鋭いピークを持ち、瞬時周波数の不動点 (=瞬時周波数と各 FFT bin の中心周波数が一致する点) とも一致していることがわかる。また、対数スペクトル上では背景雑音に由来する調波成分以外のピークが表れているが、占有度においてはこれらのピークは消失している。一方、無声区間ではピークの位置は不規則かつ不明確であり不動点の位置とも一致しない。占有度のこれらの性質は、入力信号中で占有的な調波成分の構造を抽出するのに極めて有効である。

3.2 V/UV 判定法

占有度による方法では、調波構造に対応する占有度のピーク値の和をとり、その大小で V/UV 判定する。各フレーム (1 msec シフト) ごとに以下の調波構造占有度を求め、メジアンフィルタ (61 サンプルポイント) で時間方向の平滑化処理を行った後に閾値処理する。

$$D_{t0}(f_0) = \sum_{l=1}^{n}\{D_{0,F}(l\cdot 2\pi f_0) - \mathrm{E}(D_0(\omega_c))\}. \quad (3)$$

ここで、l は高調波の次数、f_0 は F_0 の推定値、$D_{0,F}(l\cdot 2\pi f_0)$ は l 次高調波の近傍の不動点における占有度 (不動点がない場合は $\mathrm{E}(D_0(\omega_c))$) を返す関数である。なお、$\mathrm{E}(D_0(\omega_c))$ は占有度のバイアスを除去する項で占有度の周波数方向の平均値を返す関数である。

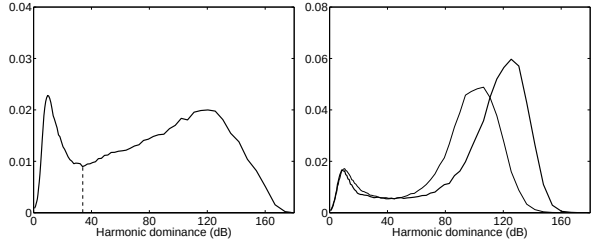

図 2: 調波構造占有度のヒストグラム (サンプル数で正規化)　左図：幼児音声データ (雑音下)、右図：成人男性 (細線) と女性 (太線) の音声 (雑音なし)

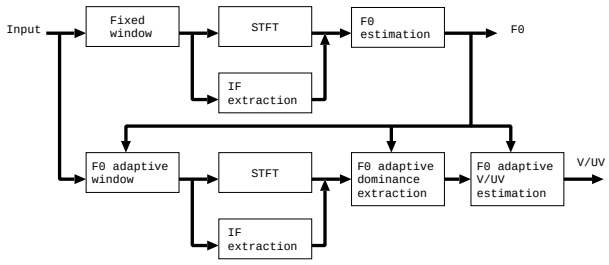

図 3: 推定のフロー

前節に示したように、有声区間では各高調波に相当する周波数と不動点および占有度のピークが一致するため、無声区間に比べて $D_{t0}(f_0)$ は大きな値をとることになる。しかも、入力音のパワーによらず、占有的な調波成分の占有度はほぼ近い値をとるため、有声区間において $D_{t0}(f_0)$ はある一定の大きな値の範囲におさまる。これに対し、無声区間の占有度は小さい値で一定の範囲におさまる。この有声/無声区間のそれぞれの占有度がとりうる範囲の境界に閾値を設定することで、V/UV 判定を行うことが出来る。

図 2 左に、幼児音声データからランダムに選んだ 1749 データから、また、図 2 右に雑音を含まない成人の音声 (28 人× 30 発話) [5] から抽出した調波構造占有度のヒストグラムを示す。各図では、横軸上の 34 dB 付近をはさんで左右に一つずつ分布の山が出来ている。左が無声区間、右が有声区間に相当する分布である。成人でも幼児でも同様な性質を持つ分布が得られ、しかも有声/無声の各分布の境界がほぼ同じ程度の値になる。これは V/UV 判定の尺度として有効であるといえる。以上の考察より、幼児音声の有声/無声の判定の閾値は、34 dB 付近に設定すればよいことがわかる。

3.3　F_0 推定法

瞬時周波数を用いた F_0 推定法 [1] のうち、幼児音声の F_0 には、占有度の代わりに残差スペクトル（パワースペクトルから包絡成分を取り除いたもの）で代替する方法（残差スペクトル法と呼ぶ）を適用する。これは、残差スペクトル法では、単純に F_0 の探索範囲を広げるだけで、広範囲の F_0 をカバーすることが出来るからである。しかも、我々が提案している F_0 推定法は、もともと倍ピッチ/半ピッチを効果的に防ぐメカニズムを備えているので、F_0 の探索範囲が広がっても、この種のエラーの増加は少ないと期待される（詳細は [1] 参照)。

3.4　推定のフロー

図 3 に、幼児音声の F_0、および V/UV 推定の処理フローを示す。まず、入力音から有声/無声区間の区別なく F_0 を推定し、この F_0 推定値を用いて V/UV 判定を行う。この方法では、瞬時周波数により高精度に推定される F_0 推定値を V/UV 判定に利用することで、1) 最適な時間窓 [5] や最適な占有度の積分範囲 [1] を用いることができるとともに、2) 正確な調波成分の周波数を特定できるので、より精度の高い V/UV 判定処理を行うことが出来ると期待される。

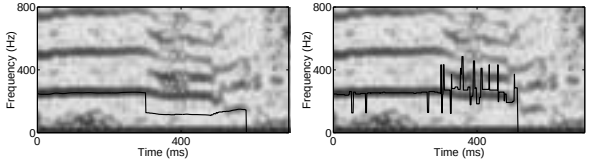

図 4: 幼児音声のスペクトルグラムと F_0 推定および V/UV の判定結果。提案法 (左)、SPTK ver. 2.0(右)

4　分析結果

提案法を用いて分析を行った結果の例を、図 4 左に示す。比較として、SPTK ver. 2.0 (以下従来法) による結果も図 4 右に示す。ここで、従来法では、F_0 の探索範囲の上限値を 2 kHz まで広げるとほとんど適切な値を返せないため、上記例では探索の上限値を 500 Hz としている。図より、提案法ではほぼ正しい F_0 および V/UV を推定できていることがわかる。特に、300 msec 付近以降にそれまでの F_0 を 1/2 倍した値の奇数倍に調波成分が現れるのに対し、F_0 が 1/2 に不連続に遷移したという適切な結果を返している。一方、従来法では、300 msec 付近で F_0 の追跡が不安定になり、正しい値が推定できなくなるとともに、500m sec 付近で有声区間が終了したと誤判定している。上記以外の幼児音声データに対しても、2 kHz まで F_0 の探索範囲を広げた場合に、提案法でほぼ良好な結果が得られることを実験により確認している。ただし、提案法が一部うまく機能しない音声データとして、1) SNR の非常に低いもの (0 dB 以下)、2) 一部の幼児音声に特徴的な調波構造が大きく乱れて非調波性を有するもの、がある。今後更なる検討が必要である。

5　まとめ

幼児音声の F_0 推定、および V/UV 判定のために、瞬時周波数に基づく占有度を用いる方法を提案した。占有度を用いることで、低 SNR 条件下でも占有的な調波成分を特定することができ、その結果、頑健に有声音の特徴抽出が行える。簡単な実験により、幼児音声に特徴的な音声の F_0 推定、V/UV 判定について提案法の有効性を示した。今後の課題として、提案法の性能の定量評価が上げられる。また、本稿ではあまり議論できなかったが、幼児音声の非調波性に関する定性的な分析が必要であろう。

参考文献

[1] 中谷他, 聴覚・音声研究会, 東大, 3 月, 2002.
[2] 天野他, ICSLP-2002, デンバー, 9 月, 2002.
[3] 河原他, *Speech Communications*, Vol. 27, Nos. 3-4 pp. 187–207, Elsevier, 1999.
[4] Ahmadi *et al.*, *IEEE Trans. SAP*, vol. 7, no. 3, pp. 333–338, 1999.
[5] 阿竹他, 信学論 Vol. J83-D-II, NO.11, pp.2077–2086, 2000.