ハフ変換による雑音に頑健な基本周波数抽出法

関高浩　　岩野公司　　古井貞熙
東京工業大学大学院　情報理工学研究科　計算工学専攻
〒152-8552 東京都目黒区大岡山 2-12-1
Email: {tseki, iwano, furui}@furui.cs.titech.ac.jp

実環境での音声認識性能の向上のため，様々な雑音環境下における頑健かつ高精度な基本周波数抽出が望まれている．本研究ではケプストラム法にハフ変換を適用し，様々な雑音に対して頑健な基本周波数抽出法を提案する．本手法では，時間-ケプストラム領域に現れる音声から得られるケプストラムのピークの軌跡を，その時間連続性を考慮することによって，ピーク位置への雑音の影響を抑えながら抽出することが可能である．4 種類の雑音を重畳させた音声を用いて，フレーム毎に抽出する従来法との性能の比較実験を行った結果，雑音の種類，大きさに関わらず，すべての場合に対して大きな抽出精度の向上がみられた．また，相関法との比較では，同等かそれ以上の精度が得られた．

Robust Pitch Extraction for Noisy Environments Using Hough Transformation

Takahiro Seki, Koji Iwano and Sadaoki Furui
Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
Email: {tseki, iwano, furui}@furui.cs.titech.ac.jp

For the improvement of speech recognition performance in real environments, robust and precise pitch (F0) extraction under various noisy environments is important. In this paper, we propose a robust pitch extraction method applying Hough transformation to the cepstrum-based pitch extraction method. This method can reduce the effect of noises on the location of cepstral peaks by tracking them taking their continuity in the time domain into account. We compared the proposed method with the conventional method which extracts pitch values frame by frame through experiments using speech data contaminated by four kinds of noise. It has been confirmed that the precision of the proposed method is much better regardless of the kind and the SNR of the noise. In addition, our method shows better performance than the correlation-based pitch extraction.

1　はじめに

これまで音声認識に関する様々な研究がなされてきた．その結果，ニュース音声などの読み上げ音声の音声認識精度はかなり高い水準にまで達している．しかし，我々の日常的な会話のような自然発話音声や，多種多様の雑音が重畳した実環境における音声に対する音声認識精度に関しては，まだ改良の余地がある．

また一方で，現在の音声認識にはイントネーションやアクセントなどの韻律情報はほとんど利用されていない．これらの韻律情報を認識性能の向上に利用しようとする研究も進められているが[1]，現状ではむしろ性能を劣化されるものとして排除される傾向にある．しかし，読み上げ音声とは違い，自然発話ではイントネーション情報がより重要な意味を持つようになると考えられる．イントネーション情報を適切にモデリングすることができれば，自然発話の音声認識精度の向上だけでなく，意味抽出や音声理解な

ど，さらに高度な音声情報処理にも役立つことが期待される．また，基本周波数の情報は従来の音声認識で用いられている音響的特徴量とは独立して抽出可能な情報であり，実環境において高精度な基本周波数抽出ができれば，雑音環境下における認識モデルの頑健性が向上するものと期待される．そこで我々は，特に後者の実環境における基本周波数情報の有用性に着目し，様々な雑音を含む音声からの頑健な基本周波数抽出法を提案する．

従来までの多くの基本周波数抽出法では，短時間窓でフレーム毎に独立に基本周波数を求めているため，特に雑音環境下ではフレーム毎に，ばらついた多数の抽出候補が出現してしまう．そのため，正しい候補の特定が困難となり，誤抽出を招く．そこで本研究では，従来法にハフ変換を適用し，ある程度長い時間における基本周波数パターンの連続性をとらえることで，様々な雑音環境に対して頑健な基本周波数抽出法を提案する．従来までの基本周波数抽出法としては主にケプストラム法，自己相関法の二つがあるが，それぞれ重畳している雑音の種類によって抽出の得意不得意が異なる．本研究ではケプストラム法にハフ変換を適用した．

2　ハフ変換

ハフ変換は雑音を含む画像から直線成分を抽出するのに有効な手法である[2]．ここでは簡単のため 2 値画像で説明する．図 1 のような 2 値画像があるとする．座標点の存在する画素が黒画素，それ以外の画素が白画素であるとする．2 値画像内の黒画素 (X_i, Y_i) だけに対して次式を用いて直線に変換する．

$$c = -X_i m + Y_i \quad (i = 1,2,\cdots,n) \quad (1)$$

ただし n は黒画素の数

つまり，図 1 に示す画像上の黒画素の各点に対して，(1)式を用いて，$m-c$ 平面(投票平面)上の軌跡を描くと図 2 に示すように，それぞれ一つの直線になる．そこで $X-Y$ 平面の各黒画素に対して，投票平面上の軌跡を描きながら，投票平面上の各点の累積度数を求める．すなわち，$X-Y$ 平面上の各黒画素に対応する投票平面の直線に相当する各点の累積度数に 1 を加えていく．そうすることで $X-Y$ 平面上での直線に対応する点の累積度数が大きい値をもつようになる．累積度数の大きな値を与える (m, c) の値によって，$X-Y$ 平面に存在する直線の傾きと Y 切片を知ることができ，直線は次式のように求めることができる．

$$Y = mX + c \quad (2)$$

図 2 の投票平面では点 $A(1, 1)$が累積度数最大の点である．点 A を(2)式を用いて直線に変換すると $Y = X + 1$となる(図 3 参照)．B 点，C 点のよ

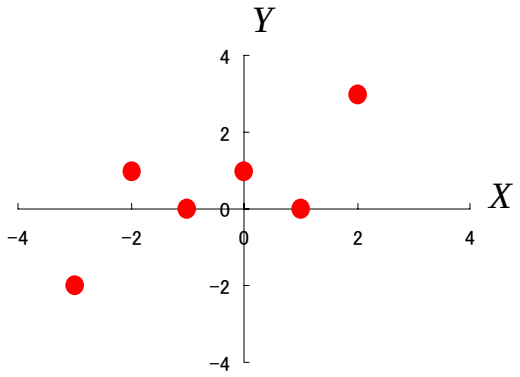


図 1　ハフ変換の対象とする 2 値画像

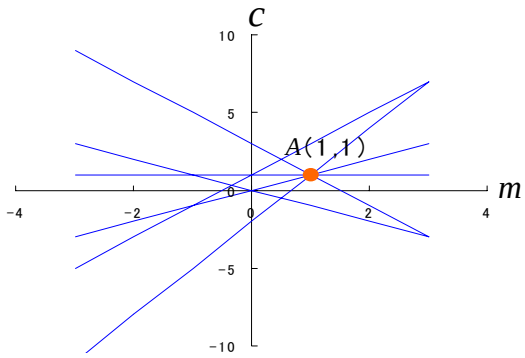


図 2　投票平面上の累積度数

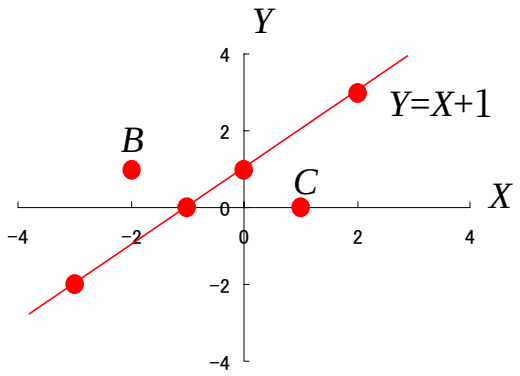


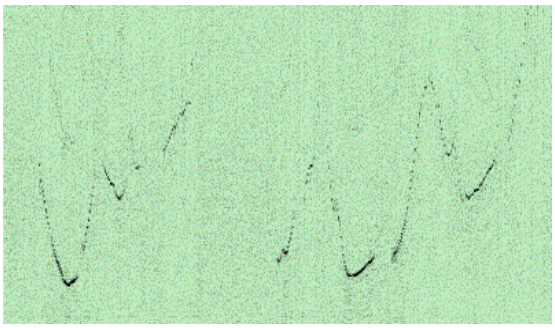
図 3　2 値画像中からの直線成分抽出

うな雑音があるにもかかわらず，$X-Y$ 平面に存在する直線が抽出されているのがわかる．

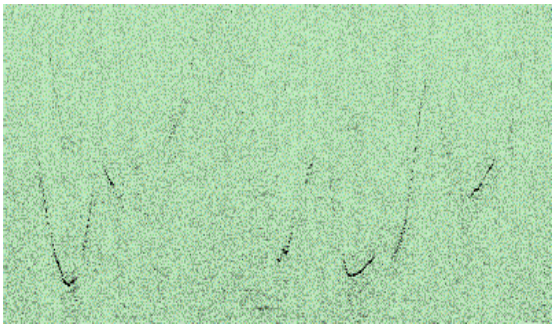
この手法は 2 値画像だけでなく，各画素が輝度値をもつ画像にも適用できる．その場合は，投票する際，直線に対応する画素の輝度値を累積度数に加えることによって，直線の抽出が実現できる．

3　ハフ変換を用いた基本周波数抽出法

本研究では，ケプストラム法にハフ変換を利用する．ケプストラム法は，各フレームについてケプストラム領域の高ケフレンシー部におけるピークを抽出し，得られたピークの次数の逆数をとることにより，そのフレームの基本周波数が決定される手法である．しかし雑音環境下では，求めるべき音声の基本周波数に対応するピークと，雑音によって発生するピークが混ざり合ってしまう．そのため，1 フレームのケプストラム情報からでは，基本周波数に対応するピークを頑健に決定できない場合が多い．図 4 に時間によるケプストラムのピーク値の推移を表す画像を示す．横軸を時間，縦軸を高ケフレンシー部におけるケプストラムの次数，各画素の輝度値をケプストラムの値とする濃淡画像である．(a)は clean な音声，(b)は SNR=10dB の雑音が重畳した音声の画像である．1 発話分で時間は約 4 秒である．(a)は比較的はっきりとピークの軌跡が描かれているが，(b)では所々に雑音によるピークが目立つ．そこで，音声の基本周波数はある程度の連続性があるという性質を利用し，このような画像に対してハフ変換を用いて，ピークの描く軌跡を頑健に抽出することで，雑音に対して頑健な基本周波数を抽出する手法を提案する．



(a)　clean な音声



(b)　雑音が重畳した音声(展示場雑音[10dB])

図 4　時間-ケプストラム次数平面における
ケプストラム濃淡画像
(発話内容「あらゆる現実をすべて自分の方へねじ曲げたのだ」)

以下に提案手法の流れを示す．

まず，音声波形から以下の音響分析条件で FFT ケプストラムに変換する．

- 標本化周波数　16kHz
- 分析窓　Hamming 窓
- 分析窓長　32ms
- フレーム周期　10ms

扱うケプストラムの次数の範囲は，男性話者の場合 60～256 次(基本周波数で約 60～270Hz に相当)，

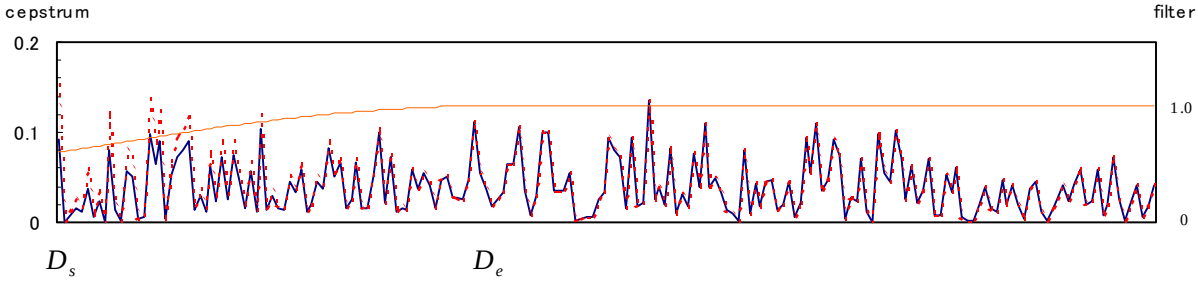


図 5　フィルタによる低次数ケプストラムの抑制
破線:フィルタ処理前のケプストラム
実線:フィルタ処理後のケプストラム

女性話者の場合30〜256次(約60〜540Hzに相当)に限定した.

次に，雑音の重畳した音声のケプストラムは，次数の低い部分ほどピーク値が大きくなる傾向があるので, 低次数のケプストラム領域でのピークを抑えるため, 上記の範囲の低次数部分に以下のフィルタを乗算する.

$$0.6+0.4\times\sin\left(\frac{d-D_s}{D_e-D_s}\times\frac{\pi}{2}\right) \qquad (3)$$

d をケプストラムの次数，D_s をフィルタ対象範囲の始点の次数，D_e を終点の次数とする．本研究では，男性話者の場合 D_s=60，D_e=140，女性話者の場合 D_s=30，D_e=140 とした．図 5 に白色雑音の重畳した男性話者の音声部分における 1 フレームのケプストラムの例を示す. 破線がフィルタリング前, 実線がフィルタリング後のケプストラムを示す.

このようにして得られたケプストラム時系列についてハフ変換を適用する. 各フレームについて, 前後 4 フレームを含む計 9 フレーム分の画像をとり出し，それを 2 節で述べた $X-Y$ 平面として，ハフ変換を行う．画素の輝度値(フィルタリング後のケプストラムの値)を投票値として投票を行い, 投票平面での累積度数が最大となる座標を求め，時間-ケプストラム平面での直線を決定する. 得られた直線の中点の Y 座標を当該フレームでのケプストラムの次数とし, 基本周波数を求める．なお，投票平面への投票については，全ての画素に対して投票を行う必要はなく, 一定以上のピーク値を持っている画素に対してのみ投票を行っても結果は変わらない．そこで，計算量削減のため，閾値 T を定め，輝度値が T 以上の画素に対して，投票平面への投票を行った．閾値 T は実験的に求め，0.05 とした.

4　評価実験

従来法であるケプストラム法とそれにハフ変換を利用した手法との比較実験を行った. 実験に使用した音声データは ATR データベースの男性話者 1 人，女性話者 1 人が発声した各 50 文である．雑音として，白色雑音，電子協雑音データベースの展示場雑音，走行車雑音，百貨店雑音の計 4 種類の雑音を用いた．重畳条件としては，それ

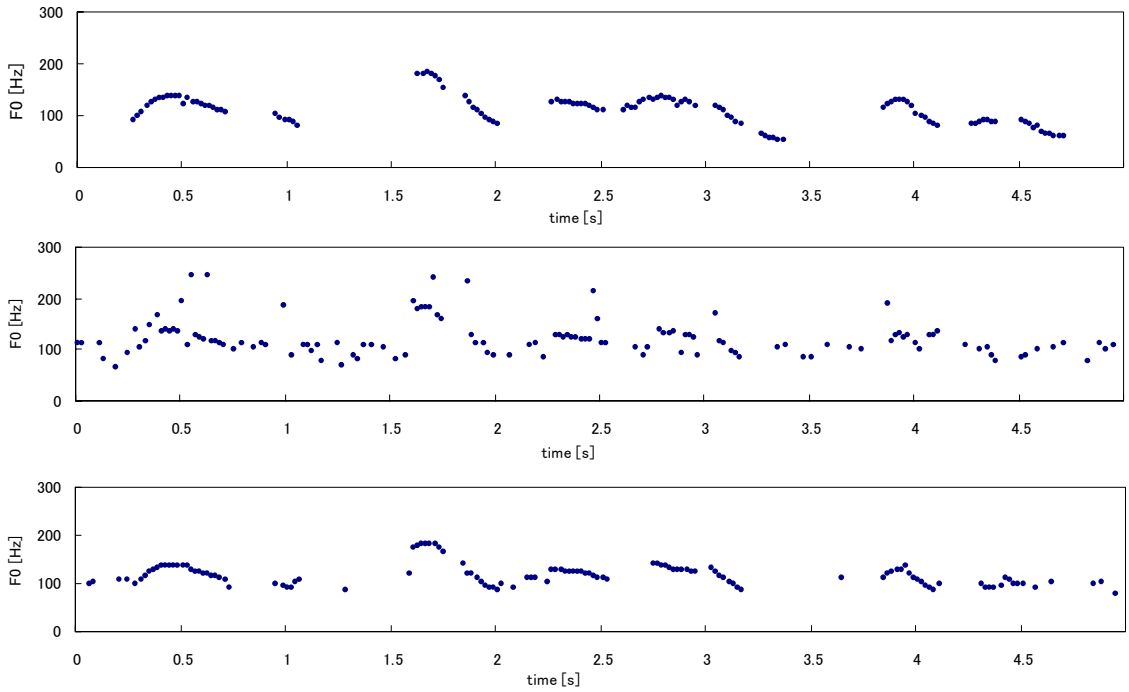


図 6: 基本周波数パターンの抽出例　「おごりを捨て謙虚な姿勢を取り戻さねば冬は過ごせない」
上段: 正解の基本周波数パターン
中段: ケプストラム法による基本周波数パターン
下段: ハフ変換を用いた手法による基本周波数パターン

ぞれの雑音に対して SNR が∞，20，10，5dB の 4 パターンで実験を行った．ケプストラム法では，当該フレームにおける高ケフレンシー部のケプストラムに(3)式でフィルタをかけ，その中で最大のピークを検出し，基本周波数を求めた．図 6 にそれぞれ正解，ケプストラム法，ハフ変換を用いた提案法による基本周波数パターンの例を示す．ここで対象とした音声は，SNR が 5dB の白色雑音を付加した男性話者の音声である．正解には ATR データベース付属の基本周波数を用いた．この際，有声/無声の判別は，ケプストラム法ではケプストラムの最大ピーク値，提案法では得られた直線に対応する投票値を閾値と比較することで行った．このとき，有声/無声を区別する閾値は，有声/無声の割合が正解のそれと等しくなるように事後的に定めた．提案法による基本周波数パターンは従来法のものに比べ，ばらつきも少なく，SNR が 5dB であるにも関わらず，かなり正解に近い基本周波数パターンが得られている．また，有声/無声の判別もより正確である．

次に定量的な比較を行う．ここでの評価にあたり，まず従来法，提案法を用いて全てのフレームで基本周波数を推定しておく．その上で，正解の有声区間と一致するフレームについて，以下の式により抽出精度を示す正解率を定義する．

$$\text{正解率} = \frac{\text{正解の基本周波数と}\pm 5\%\text{以内の誤差となるフレーム数}}{\text{正解の有声部分のフレーム数}}$$

また，参考のために Entropic 社の音声分析ツール ESPS の基本周波数抽出プログラム get_f0[3] に対しても同様の評価実験を行った．get_f0 は相互相関関数を用いた手法で，自己相関法の一種である．白色雑音，展示場雑音，走行車雑音，百貨店雑音を付加した女性話者の音声に対する正解率を図 7，8，9，10 に示す．ケプストラム法を cepstrum，ハフ変換を利用した手法を hough，ESPS の get_f0 を get_f0 で示す．すべての雑音で SNR の大きさに関わらず，ハフ変換を用いた提案法が従来法であるケプストラム法よりも推定精度が大きく向上しているのがわかる．特に SNR が小さいときほど正解精度が向上する傾向が見られた．SNR が 5dB の場合，白色雑音を付加した場合では 20.2%，展示場雑音では 13.4%，走行車

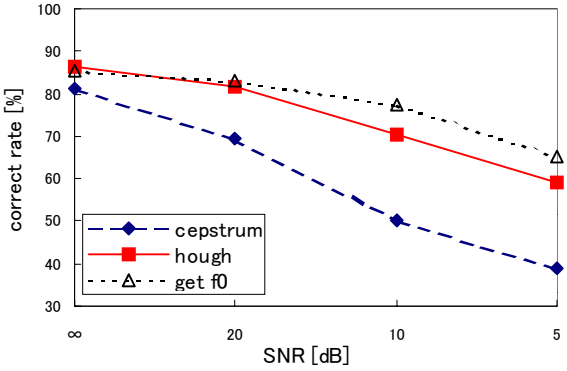


図 7　白色雑音を付加した女性音声に対する正解率

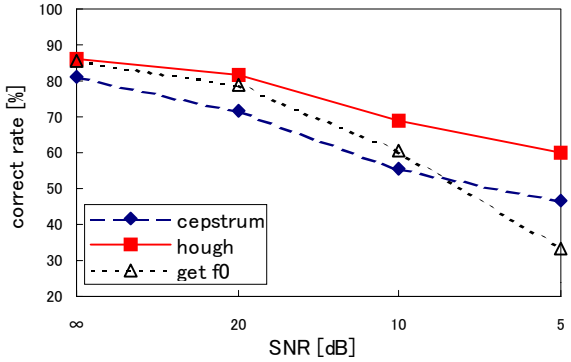


図 8　展示場雑音を付加した女性音声に対する正解率

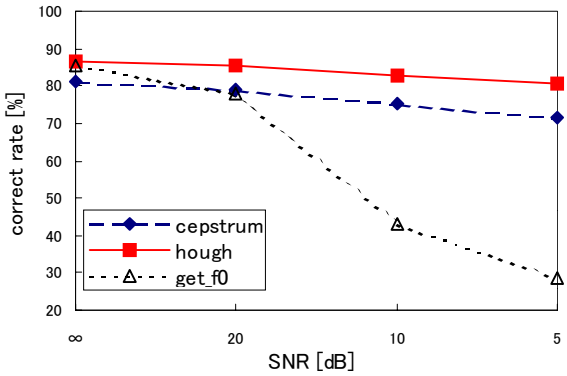


図 9　走行車雑音を付加した女性音声に対する正解率

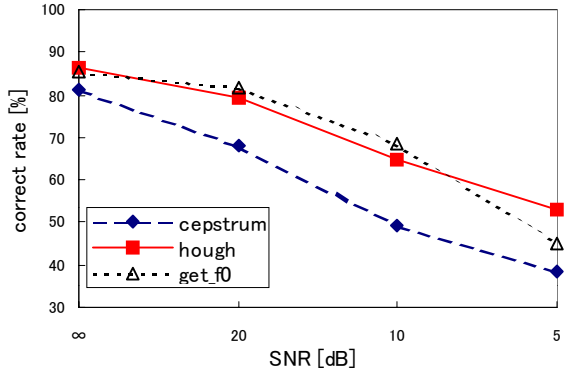


図 10　百貨店雑音を付加した女性音声に対する正解率

雑音では 9.1%, 百貨店雑音では 14.6%の正解率の向上が示された．また，一般的に白色雑音のような周期性のない雑音に対しては, ケプストラム法よりも自己相関法の方が抽出精度が高いが, ハフ変換を用いることによって get_f0 と同等の精度が得られている．また，展示場，走行車，百貨店雑音のような一般的に雑音に関しては，get_f0 は SNR が小さくなるほど性能の劣化が激しいのが見てとれるが, そのような場合でも提案法は高い抽出精度を維持しているのがわかる. 男性話者の音声に対しても同様の結果が得られ，SNR が 5dB の場合, 白色雑音を付加した場合では 18.2%, 展示場雑音では 11.8%, 走行車雑音では 7.0%, 百貨店雑音では 11.1%の正解率の向上が確認された.

5 まとめ

ケプストラム法にハフ変換を用いて前後フレームにおけるケプストラムのピークの連続性をとらえ，軌跡を直線成分として抽出することで，雑音環境下において頑健かつ高精度な基本周波数を抽出する手法を提案した. ケプストラム法との抽出精度比較実験を行い，雑音の種類，大きさに関わらず, 全ての場合において提案手法の有効性が示された．また，提案手法は相関法である get_f0 と比較すると，一般的な雑音に対してはより高い精度で, ケプストラム法の不得意とする雑音に対しては get_f0 と同等の精度で抽出が実現できた．したがって，様々な雑音に対して提案手法の有効性が示された.

本研究ではケプストラム法にハフ変換を利用した. 他の基本周波数抽出法の多くはケプストラム法同様，ピークの検出によるものであるが，時間によるピークの推移を画像として表現することができれば, これらの手法においてもハフ変換を適用することができる. したがって自己相関法などにも応用が可能である.

本手法は，計算時間が非常に大きいため，実時間での利用を考えると, 計算量の削減が課題となる．また，今後は本手法で推定した基本周波数情報を利用することによって, 雑音に対して頑健な音声認識手法の提案を目指す.

参考文献

[1] Y. Sagisaka, N. Campbell, and N.Higuchi, eds., Computing Prosody, Part Ⅳ, Springer-Verlag, New York, 1997.

[2] 安居院猛，中島正之，画像情報処理，森北出版, pp. 115-117, 1999.

[3] Entropic research laboratories, esps/waves+ with ensig5.3, reference release, 1998.