# 3—1—5　自発的な発話中の言い淀み (有声休止) 箇所のリアルタイム検出*

◎ 後藤 真孝　伊藤 克亘　速水 悟　(電総研)

## 1. はじめに

自然な発話には，有声休止，無声休止，音節の引き延ばし，言い直しといった，話し言葉特有の言い淀み現象が頻繁に現れる[1),2)]．音声対話システムの性能を向上させるには，言い淀み現象を冗長語や不用語等とみなして単に無視するのではなく，言い淀みが起きていることを的確に認識し，それらの役割を把握して活用することが重要である．その第一段階として，本稿では，音声対話において共通した大切な役割を果たしている，**有声休止** (filled pause) と **音節の引き延ばし** (word lengthening) の二つを取り上げる．

典型的な音声認識システムは，言い淀み現象を含まない朗読音声を前提としてきたため，自然な発話の認識は困難である．そこで従来，連続音声認識やワードスポッティングの枠組みで有声休止を部分的に扱う手法が提案されてきた[3)~5)]．しかしこれらは，言い淀みを個々に検出し，その役割を把握して扱うアプローチではなかった．一方，音響分析による検出可能性については，文献6),7) において示唆されていたが，韻律的特徴の調査に留まっており，有声休止を自動検出するシステムはまだ構築されていなかった．

本稿では，自然な発話による音響信号に対して，有声休止と音節の引き延ばしの二つの言い淀み現象を，ボトムアップな音響分析によって検出する手法を提案する．両者は同様な音響的特徴を持っており，対話の観点からは同じ機能を果たしていると考えられるため，以下,「有声休止」を両者を指す用語として用いる．

## 2. 有声休止の重要性

本研究では，有声休止が自然な発話において不可避なのは，それが，思考プロセスが発話プロセスに追い付かない場合に表れる現象であるからだと考える．次の発話内容が発話プロセスに届くまでの間，話者は，時間を稼ぐために有声休止や無声休止を用いる．

有声休止の検出は，大別して二つの意義を持つ．一つは音声認識に対する貢献で，例えば，検出した有声休止区間を考慮して認識することで，音声認識性能の向上が期待できる．もう一つは音声対話に対する貢献で，有声休止の役割を考慮した音声対話システムを実現することが可能になる．有声休止は，対話において，少なくとも次の二つの大切な役割を担っている[1),8),9)]．

- 発話権の保持，場つなぎの機能
  話者は，次の発話が準備できていないにも関わらず発話権を持ち続けたいとき (あるいは何か発話しなければならない状況のとき)，有声休止によって，聴取者に次の発話を待って欲しいと伝えることができる．聴取者は，有声休止を聞くと，話者の次の発話を待った方がよい等と判断できる．
- 話者の心的状態・思考状態を表す機能
  話者は，有声休止の発声法等によって，発話内容に対する自信のなさ，不安，躊躇，謙遜といった心的状態を表現できる．また，そのつなぎ語の種類等によって，異なる思考状態も表現できる．聴取者は，有声休止から話者の現在の心的状態・思考状態を推測し，それを言語以外の付加情報 (別のモダリティ) として利用できる．さらに，次の発話内容を予測することも，場合によっては可能となる．

## 3. 有声休止検出手法

本手法では，有声休止の音響的特徴を，ボトムアップな信号処理で検出する．次の発話内容が間に合わないときに有声休止が発声されるのであれば，話者は調音器官の位置・状態を，有声休止中に変化させることができない．そこで有声休止は，調音器官がほぼ一定のまま声帯が振動し続けるときの音声，つまり，音韻的に変化が少ない持続した有声音 (以下，**有声休止音**) を伴っていると仮定する．実際に，典型的に用いられるつなぎ語 (「えー」「あのー」等) や音節中の母音の引き延ばし箇所には，有声休止音が含まれている．

以上から本手法では，有声休止音が持つ次の二つの特徴に基づいて，有声休止を検出する．

1. 基本周波数の変動が小さい．
   声帯の緊張条件が変化しなければ，声の基本周波数はほぼ一定となる．
2. スペクトル包絡の変形が小さい．
   声道形状が変化しなければ，フォルマントを反映したスペクトル包絡はほぼ一定となる．ただし，肺からの呼気量は変化しうるため，そのAM変調成分を取り除いて，スペクトル包絡の変形量を評価する必要がある．

以下，図 1に沿って処理の流れを順に説明する．

### 3.1 瞬時周波数の算出と周波数成分の抽出

まず，Flanaganの手法[10)] を用い，STFTの出力をフィルタバンク出力と解釈して，効率良く瞬時周波数を計算する．本実装では，16 kHz / 16 bitでA/D変換し，1024点のハニング窓を用いたSTFTを160点 (10 msec) ずつフレームシフトする．次に，STFTフィルタの中心周波数からその出力の瞬時周波数への写像の安定平衡点を，周波数成分として抽出する[11)~13)]．これらの各瞬時周波数におけるSTFTパワースペクトルの値を，周波数成分のパワー分布関数として求める．

### 3.2 基本周波数の推定

背景雑音・音楽を伴う実世界の音響信号に対してロバストに機能するように，LPC等の単一音源を前提とした分析はおこなわず，入力信号中で最も優勢な (パワーの大きい) 高調波構造の基本周波数を，音声の基本周波数として抽出する．そこで，コムフィルタの考

---

* "Real-time Detection of Filled Pauses in Spontaneous Speech" by M. Goto, K. Itou and S. Hayamizu (ETL)

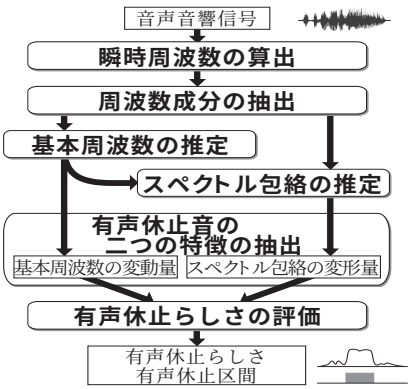


図 1: 有声休止検出手法の処理の流れ

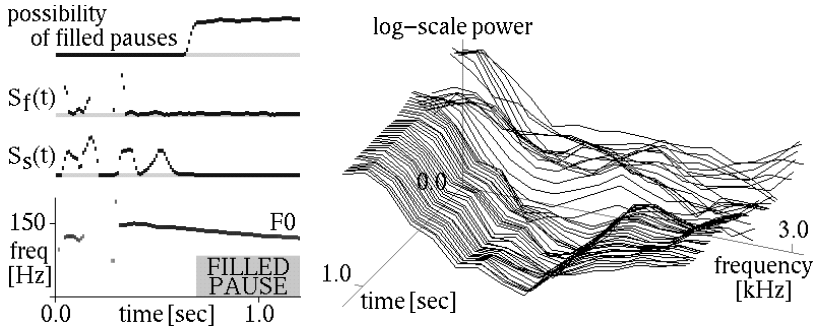


図 2: システムの画面表示例: 男性の自由発話の一部である「一階にー」/iqkaini-/ を入力し終わった時点での，基本周波数とその変動量，スペクトル包絡の変形量，有声休止らしさ，有声休止区間 (左側)，および対応するスペクトル包絡 (右側)

え方に基づいて，時刻 $t$ において周波数 $F$ が基本周波数となる可能性 $P_{F0}(F,t) = \int_{-\infty}^{\infty} p(x;F)\ \Psi_p(x,t)\ dx$ を評価する．$p(x;F)$ は基本周波数が $F$ の高調波成分を通過させるフィルタ関数，$\Psi_p(x,t)$ は周波数成分のパワー分布関数とする．$P_{F0}(F,t)$ は各高調波構造が相対的にどれくらい優勢かを表すため，基本周波数 $F_{F0}(t)$ は，$F_{F0}(t) = \mathrm{argmax}_F\ P_{F0}(F,t)$ で求まる．

### 3.3 スペクトル包絡の推定

実環境でロバストに推定するために，$F_{F0}(t)$ の高調波構造上にある局所的な情報だけを利用する．まず，各高調波成分のパワーを，基本周波数の整数倍を中心とするガウス分布で重み付けしながら，その近傍の最大パワーを検出することで求める．次に，隣接する成分のパワーの間を直線補間してスペクトル包絡を求める．包絡の計算では，日本語の母音の第一，第二フォルマントを捉えられるような上限周波数 (3200 Hz) を設ける．有声休止音の特徴としては，包絡の大局的な変形を捉えた方が良いため，直線補間した包絡を粗い周波数分解能 (200 Hz) でリサンプリングし，低い方から $n$ 点目の周波数におけるスペクトル包絡 $Env(n,t)$ を求める．最後に，肺からの呼気による AM 変調の影響を除去できるように，$Env(n,t)$ を正規化する．

### 3.4 有声休止音の二つの特徴の抽出

有声休止音の二つの特徴として，基本周波数の変動量 $A_f(t)$ とスペクトル包絡の変形量 $A_s(t)$ を求める．$A_f(t)$ は，対数スケールの基本周波数 $F_{F0}(t)$ の過去一定期間 (50 ms) の変化を，最小自乗法で直線近似した直線の傾き $b_{F0}$ を用いて，$A_f(t) = \mid b_{F0} \mid$ のように定義する．一方，$A_s(t)$ は，スペクトル包絡 $Env(n,t)$ の対数スケールのパワーの過去一定期間 (100 ms) の変化を，最小自乗法で直線近似した際の直線の傾き $b_s(n)$ と誤差 $err_s(n)$ を用いて，$A_s(t) = \left(\frac{1}{\mathrm{N_{max}}}\sum_{n=1}^{\mathrm{N_{max}}} b_s(n)^2\right)\left(\frac{1}{\mathrm{N_{max}}}\sum_{n=1}^{\mathrm{N_{max}}} err_s(n)^2\right)$ のように定義する ($\mathrm{N_{max}} = 15$)．

### 3.5 有声休止らしさの評価

**有声休止らしさ** (有声休止と判定する信頼度) $P_{fp}(t)$ は，$A_f(t)$，$A_s(t)$ の短時間 (100 ms) 平均 $S_f(t)$，$S_s(t)$ に基づいて，$P_{fp}(t) = \exp\left(-\frac{(\mathrm{R}\ S_f(t) + (1-\mathrm{R})\ S_s(t))^2}{\mathrm{W}^2}\right)$ のように定義する．R は特徴に対する重み付けを決める定数，W は変動・変形の考慮範囲を決める定数とする．最終的に，$P_{fp}(t)$ が一定期間十分高い値のときに，話者が有声休止をおこなっていると判定する．

## 4. 実験結果

音声音響信号を入力し，有声休止らしさと有声休止区間の判定結果をリアルタイムに出力するシステムを，提案手法に基づいて構築した (図 2)．日本語の自由発声音声の対話コーパス[14] から，100 発話 (男女 5 名ずつ，各話者当り有声休止を最低一つ含む 10 発話) を抜粋して実験した．実験の結果，再現率 (正しく検出した数 / 有声休止の総数) は 84.9 % (107 / 126)，適合率 (正しく検出した数 / 検出した総数) は 91.5 % (107 / 117) であった．検出もれは，基本周波数の変化が大きすぎる箇所や，声がしわがれて高調波成分が乱れた箇所で起きていた．誤検出は，平坦な基本周波数で発声された，変化の少ない持続した有声音で起きていた．

## 5. おわりに

本稿では，音韻的に変化が少ない持続した有声音を見つけることで，音韻やつなぎ語の種類を問わずに，有声休止 (音節の引き延ばしも含む) の箇所を検出する手法について述べた．なお本稿では割愛したが，有声休止を含む発話に対する音声認識で，本手法を用いることによる性能向上 (アラインメントの改善) が示唆されている[15),16)]．今後は有声休止の役割を積極的に活用した音声対話システムを構築していく予定である．

### 参 考 文 献


[1] 田窪: 音声言語の言..., 情処学会誌, **36**(11), 1020–1026, 1995.
[2] 伊藤: 音声対話システム, 自然言語処理, 信学会, 302–322, 1999.
[3] Ward: Understanding Spon..., *ICASSP 91*, 365–367, 1991.
[4] 中川 他: 自然な音声対話にお..., 音響誌, **51**(3), 202–210, 1995.
[5] 甲斐 他: 冗長語..., 信学論, **J80-D-II**(10), 2615–2625, 1997.
[6] O'Shaughnessy: Recognit..., *ICASSP 92*, I–521–524, 1992.
[7] Quimbo *et al.*: Prosodic analysis of fill..., *ICSLP 98*, 1998.
[8] 田中: 「休止」の意味論, 言語, **22**(8), 20–27, 1993.
[9] Rose: *The communicative value of filled pauses in spontaneous speech*, Master's thesis, Univ. of Birmingham, 1998.
[10] Flanagan *et al.*: Phase Vocoder, *The Bell System Technical J.*, **45**, 1493–1509, 1966.
[11] Charpentier: Pitch detect..., *ICASSP 86*, 113–116, 1986.
[12] 阿部 他: 瞬時周..., 信学論, **J79-D-II**(11), 1771–1781, 1996.
[13] 河原 他: 瞬時周波数を用いた基..., 音楽音響研 H-98-116, 1998.
[14] Itou *et al.*: A Japanese spontan..., *JASJ (E)*, **20**(3), 1999.
[15] 後藤 他: 自然発話中の言い淀み箇所のリアルタイム検出システム, 情処研報 音声言語情報処理研究会, 99-SLP-27-2, 1999.
[16] Goto *et al.*: A Real-time Filled Pause Detection System for Spontaneous Speech Recognition, *Eurospeech 99*, 1999.