音声言語情報処理 27－2
（1999．7．23）

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

後藤 真孝　伊藤 克亘　速水 悟

電子技術総合研究所
{goto, kito, hayamizu}@etl.go.jp

あらまし　本稿では，代表的な言い淀み現象である有声休止(音節の引き延ばしも含む)を自動的に検出する手法を提案する．有声休止は音声対話において発話権の保持等の大切な役割を果たしており，その検出は音声対話システムを実現する上で重要である．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に対処する研究事例はあったが，言い淀み現象として個々に検出しておらず，その役割を把握して適切に扱うことはできなかった．本手法は，有声休止中は調音器官の変化が小さいことを仮定し，基本周波数の変動とスペクトル包絡の変形が小さい箇所を有声休止として検出する．日本語の音声対話コーパスを用いた実験の結果，再現率84.9％，適合率91.5％の精度で検出できることが確認された．

A Real-time System Detecting Filled Pauses in Spontaneous Speech

Masataka Goto　Katunobu Itou　Satoru Hayamizu

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki 305-8568 Japan

Abstract　This paper describes a method for detecting filled pauses (including word lengthening), which are one of the hesitation phenomena. This detection is important in speech dialogue systems because they play valuable roles in oral communication. Although there have been a few previous speech recognition systems handling filled pauses, they have not detected them individually and consequently could not consider their roles. Our method can detect filled pauses by finding small fundamental frequency transition and small spectral envelope deformation under the assumption that articulator parameters do not change during filled pauses. Experimental results for a Japanese spoken dialogue corpus show that our system yielded a recall rate of 84.9% and a precision rate of 91.5%.

1　はじめに

本研究の最終的な目標は，計算機の音声理解能力を向上させ，人間と計算機との間で自然なマルチモーダル対話を実現することである．そのためには，話者がその場で内容を考えながら自発的に発話した音声を，計算機が理解できる必要がある．そのような自然な発話には，有声休止，無声休止，音節の引き延ばし，言い直しといった，書き言葉には通常現れない，話し言葉特有の言い淀み現象が頻繁に現れる[1),2)]．このような自然で不可避な現象を扱うための第一段階として，本稿では，有声休止(filled pause)と音節の引き延ばし(word lengthening)の二つの言い淀み現象を対象に議論する．この二つを取り上げたのは，音声対話において，これらが共通して，発話権の保持や心的状態・思考状態の表出といった大切な役割を果たしているからである．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と我々は考えている．

典型的な音声認識システムは，言い淀み現象を含まない，書き言葉を読み上げたような朗読音声を前提としてきたため，自然な発話を認識することは一般に困難である．例えば，HMMに基づく音韻モデルを，有声休止や音節の引き延ばしを伴う音声に適用すると，音韻の継続時間が突然大きく延びることがあるため，有効に機能しなくなる．また，言語モデルに関

しても，有声休止はほとんど任意の単語間に入りうるため，それを網羅的に記述したような文法は，制約としては弱くなってしまい効果的でない．そこでこれまでに，このように誤認識の原因となる有声休止を，サブワード単位に基づく連続音声認識やワードスポッティングの枠組みで部分的に扱う手法が提案されてきた[3)~6)]．例えば，10個のつなぎ語☆1を語彙に追加登録することによって連続音声認識システムで扱えるようにする手法[4)]や，つなぎ語を未知語とみなして，サブワード系列照合に基づく未知語処理で対処する手法[5),6)]等が既に提案されている．しかしこれらは，言い淀み現象を個々に検出し，その役割まで把握しながら適切に扱うようなアプローチではなかった．

そこで本研究では，有声休止と音節の引き延ばしの箇所を，ボトムアップな音響分析によって個々に検出するアプローチをとる．音響分析による検出の実現可能性については，つなぎ語の韻律的特徴に関する従来研究[7),8)]において既に示唆されている．特に，文献8)は，人間は馴染みのない外国語に対しても，韻律的な手がかりから言い淀みが検出できることを指摘し，ボトムアップに韻律的特徴を分析するアプローチを支持している．しかし，これらの研究は韻律的特徴の調査に留まっており，自動的に有声休止を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検出する手法を提案する．両者の言い淀み現象は同様な音響的特徴を持っており，音声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，本稿では以下，「有声休止」を両者を指す用語として用いる．以下の章では，まず，有声休止の役割を考察し，提案手法のアルゴリズムを説明する．次に，その手法に基づいて有声休止をリアルタイムに検出するシステムの実装について述べ，それを用いた実験結果を示す．最後に，音声認識の枠組みに本手法を導入することで，認識率を向上できる可能性があることを検討する．

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本質的に不可避なのは，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．その場で内容を思考しながら発話する場合，発話スピードとその内容を準備する思考スピードとは必ずしも一致しない．そこで，思考スピードの方が遅い場合(そもそも思考対象が何かわからない場合等も含む)，思考プロセスの結果であ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．

音響信号中の有声休止の区間を検出することは，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，検出した有声休止の区間を除いてから認識処理をおこなうことで，自然発話に対する音声認識システムの性能を向上させられることが期待できる．これについては，6章で改めて議論する．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，文献1),9),10)等でも述べられているように，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と考えられる．

- 発話権の保持，場つなぎの機能

  音声対話では，その進行に伴い，話者間で発話権が移動していく．話者の立場からは，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(あるいはとりあえず何か発話しなければならない状況のとき)，発話を準備しながら有声休止をおこなうことで，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．逆に聴取者の立場からは，有声休止を聞くと，割り込んで発話権を奪うのを控え，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

  円滑な対話を進めるために，話者は自分の心的状態・思考状態を，無意識のうちに聴取者と共有する行動をとる．話者の立場からは，有声休止の方法(音韻やイントネーション，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，何かを思い出そうとしているのか，あるいは聴取者にとって適切な表現を探しているのかといった，異なる思考状態を表現することができる(例えば「ええと」と「あのー」の使い分け等が文献1)で議論されている)．逆に聴取者の立場からは，有声休止を解釈することで，話者の現在の心的状態・思考状態を推測することができ，それを言語情報以外の付加情報(別のモダリティ)として利用できる．さらに，次の発話内容をある程度予測することも，場合によっては可能となる．その際に

☆1 多くの論文では，言い淀む際に用いる「えーと」や「あのー」等の語を間投詞と呼んでいるが，場つなぎ的な役割をより明確に表すために，本稿ではつなぎ語(filler)という用語を用いる．

は，前述のように話者の発話を待つだけでなく，話者の手助けとなるような発話をおこなうこともある．

我々は特に，有声休止を積極的に活用した音声対話システムを構築することを目指している．その第一段階として，例えば，ユーザが有声休止をしていることを音声対話システム側が検出したとき，システムに次のような対応をとらせることを検討している．

- 検出した有声休止の区間では，ユーザの思考を妨げないように，システム側は相槌を打ったり確認発話をしたりせずに次の発話を待つ．
- 有声休止を検出した時点で，システム側がユーザの次の発話を予測できれば，その予測内容を提示してユーザの発話の手助けをする．そのためには，次の発話の予測を常におこなうような仕組みを導入する必要がある．スロットを埋めるようなタスクの場合には，スロットに入りうる候補を提示するのもよい．
- 文献11)で提案された，休止を区切りとした対話処理において，提案時には無声休止しか言及されていなかったが，有声休止(特に，音節の引き延ばし)箇所も区切りの候補として利用する．これを間接的に支持する知見として，対話でなく講演調の話し言葉に対する分析ではあるが，つなぎ語の箇所が知覚的な区切りとなることが報告されている[12)]．

逆に，システム側の発話に関しても，前述の機能を効果的に使う目的で，有声休止を導入していく予定である．

3 有声休止検出手法

本手法では，音声音響信号中の有声休止の音響的特徴を，ボトムアップな周波数解析によって検出する．2章で述べたように，発話プロセスが思考プロセスから次の発話内容が届くのを待っている間に，有声休止が発声されるのであれば，話者は調音器官(喉頭を含む)の位置・状態を有声休止中は変化させることができない．調音器官をどう動かすかを決めるはずの次の発話内容が，まだ準備されていないからである．そこで本手法では，有声休止は，調音器官がほぼ一定のまま(声道形状がほとんど変化しない状態で)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，有声休止で典型的に用いられる「えー」「うー(ん)」「あー」「まー」「んー」「あのー」「そのー」「このー」等や，音節中の母音の引き延ばし箇所には，このような有声休止音が含まれており，これが妥当な仮定であることがわかる．

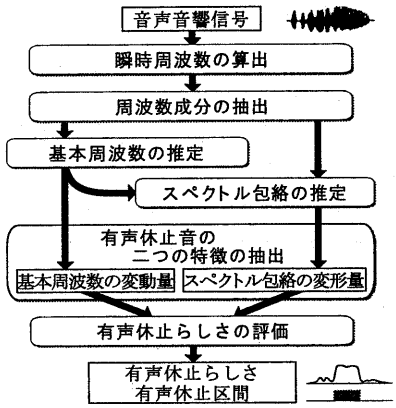

図 1: 有声休止検出手法の処理の流れ

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．

   調音器官の状態が一定であれば，声帯の緊張条件は変化せず，声の基本周波数はほぼ一定のままとなる．

2. スペクトル包絡の変形が小さい．

   調音器官の位置が一定であれば，声道形状は変化せず，フォルマントを反映したスペクトル包絡はほぼ一定のままとなる．ただし，有声休止中でも肺からの呼気量は変化するため，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．

我々の提案する有声休止検出手法の処理の流れを図1に示す．まず，入力音響信号に対して瞬時周波数を計算し，瞬時周波数に関連した尺度に基づいて周波数成分を抽出する．次に，基本周波数を推定し，その結果に基づいてスペクトル包絡を推定する．その際，背景雑音や背景音楽を伴う入力に対しても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，最も優勢な高調波構造に基づく推定をおこなう．そして，前述した有声休止音の二つの特徴を定量的に捉え，それらを統合して，有声休止であると判定する信頼度「有声休止らしさ」を評価する．最後に，有声休止らしさとそれに基づいて決定した有声休止区間を出力する．

3.1 瞬時周波数の算出

本手法では，まず，フィルタバンクの各出力信号に対し，位相の時間微分である瞬時周波数[13),14)]を計算する．ここでは，Flanaganの手法[13)]を用い，短時間フーリエ変換(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．入力音響信号 $x(t)$ に対する窓関数 $h(t)$ のSTFTが

$$X(\omega,t)=\int_{-\infty}^{\infty}x(\tau)h(\tau-t)e^{-j\omega\tau}d\tau \quad (1)$$

$$=a+jb \quad (2)$$

で定義されるとき，瞬時周波数 $\lambda(\omega,t)$ は次式で求めることができる．

$$\lambda(\omega,t)=\omega+\frac{a\frac{\partial b}{\partial t}-b\frac{\partial a}{\partial t}}{a^2+b^2} \quad (3)$$

現在の実装では，音響信号を標本化周波数 16 kHz，量子化ビット数 16 bitでA/D変換し，窓関数 $h(t)$ として窓幅1024点のハニング窓を用いたSTFTを，高速フーリエ変換(FFT)によって計算する．その際，FFTのフレームを160点ずつシフトするため，フレームシフト時間(1フレームシフト)は 10 msec となる．このフレームシフトを，すべての処理の時間単位とする．

3.2 周波数成分の抽出

フィルタの中心周波数からその瞬時周波数への写像に基づいて，周波数成分を抽出する[15)~17)]．あるSTFTフィルタの中心周波数 ω からその出力の瞬時周波数 $\lambda(\omega,t)$ への写像を考える．すると，もし周波数 ψ の周波数成分があるときには，ψ がこの写像の不動点に位置し，その周辺の瞬時周波数の値はほぼ一定となる[17)]．つまり，全周波数成分の瞬時周波数 $\Psi_f(t)$ は，次式によって抽出することができる[18)]．

$$\Psi_f(t)=\{\,\psi \mid \lambda(\psi,t)-\psi=0,\ \frac{\partial}{\partial\psi}(\lambda(\psi,t)-\psi)<0\} \quad (4)$$

これらの周波数成分のパワーは，$\Psi_f(t)$ の各周波数におけるSTFTパワースペクトルの値として得られるため，周波数成分のパワー分布関数 $\Psi_p(\omega,t)$ を次のように定義できる．

$$\Psi_p(\omega,t)=\begin{cases}|X(\omega,t)| & \text{if } \omega\in\Psi_f(t)\\ 0 & \text{otherwise}\end{cases} \quad (5)$$

3.3 基本周波数の推定

抽出した周波数成分に基づいて，話者の音声の基本周波数を推定する．その際，実験環境における話者単独の音声だけでなく，背景雑音や背景音楽を伴うような，実世界の音響信号中の話者の音声にも適用できるようにすべきである．そこで，非周期的な雑音に加え，高調波構造を持つ弱い雑音も含まれる場合を考慮して，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．そのために，コムフィルタの考え方に基づいたフィルタを用いて，時刻 t において周波数 F が基本周波数となる可能性 $P_{F0}(F,t)$ を評価する．なお，本稿では以下，対数スケールの周波数をcentの単位(本来は音高差(音程)を表す尺度)で表し，Hzで表された周波数 $f_{\rm Hz}$ を，次のようにcentで表された周波数 $f_{\rm cent}$ に変換する．

$$f_{\rm cent}=1200\log_2\frac{f_{\rm Hz}}{{\rm REF}_{\rm Hz}} \quad (6)$$

$${\rm REF}_{\rm Hz}=440\times 2^{\frac{3}{12}-5} \quad (7)$$

基本周波数の可能性 $P_{F0}(F,t)$ は，次式のように定義する．

$$P_{F0}(F,t)=\int_{-\infty}^{\infty}p(x;F)\,\Psi'_p(x,t)\,dx \quad (8)$$

ここで，周波数を表す x と F の単位はcentとし，$p(x;F)$ は基本周波数が F の高調波成分だけを通過させるフィルタ関数，$\Psi'_p(x,t)$ は，周波数軸がcentで表されていることを除けば $\Psi_p(\omega,t)$ (式 (5)) と同じパワー分布関数であるとする．フィルタ関数 $p(x;F)$ は次式のように与える(図 2)．

$$p(x;F)=\sum_{h=1}^{\rm N}c(h)\,G(x;F+1200\log_2 h,{\rm W_f}) \quad (9)$$

$$G(x;m,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \quad (10)$$

ここで，N (現在の実装では8) は考慮する高調波成分の数(基本周波数成分も数える)，$\rm W_f$ (20 cent) はガウス分布 $G(x;m,\sigma)$ の標準偏差を表す．$c(h)$ は，第 h 次高調波成分の通過量を決める関数で，本研究では $c(h)={\rm H}^{h(h-1)}$ ($\rm H=0.9849$) とする．

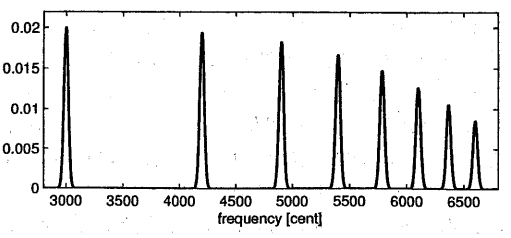

図 2: 基本周波数 F の高調波成分だけを通過させるフィルタ関数 $p(x;F)$ ($F=3000$ centの場合を図示)

こうして求めた $P_{F0}(F,t)$ は，各高調波構造が相対的にどれくらい優勢かを表しているため，話者の音声の基本周波数 $F_{F0}(t)$ は，$P_{F0}(F,t)$ を最大にする周波数として求めることができる．

$$F_{F0}(t) = \underset{F}{\operatorname{argmax}}\ P_{F0}(F,t) \tag{11}$$

3.4 スペクトル包絡の推定

実環境でロバストにスペクトル包絡を推定するために，得られた基本周波数 $F_{F0}(t)$ 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，$F_{F0}(t)$ の第 k 次高調波成分のパワー $Pow(k,t;F_{F0}(t))$ を，基本周波数の整数倍の周波数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．

$$Pow(k,t;F_{F0}(t)) = \max_{x} G(x;F_{F0}(t)+1200\log_2 k, \mathrm{W_s})\ \Psi'_p(x,t) \tag{12}$$

ここで，$\mathrm{W_s}$ (35 cent) はガウス分布の標準偏差を表す．

次に，線形スケールの周波数軸上で，隣接する $Pow(k,t;F_{F0}(t))$ の間を直線補間して，スペクトル包絡を求める．この包絡の計算は，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) を設けておこなう．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 ξ (200 Hz) でリサンプリングし，低い方から n ($1 \le n \le \mathrm{N_{max}}$ (15)) 点目の周波数 $n\xi$ 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 変調の影響を除去するために，条件

$$\sum_{n=1}^{\mathrm{N_{max}}} Env(n,t) = 1 \tag{13}$$

を満たすように $Env(n,t)$ を正規化する．

3.5 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．前者は，基本周波数の変動がどれくらい大きいかを表し，後者は，スペクトル包絡の変形がどれくらい大きく，一様でないかを表す．

基本周波数の変動量 $A_f(t)$ 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，次式のように定義する．

$$A_f(t) = \left| b_{F0} \right| \tag{14}$$

b_{F0} は，a_{F0} と b_{F0} をパラメータとして次式を最小化することで得られる．

$$err_{F0}{}^2 = \sum_{\tau=0}^{\mathrm{Period_{F0}}-1} \Big(F_{F0}(t-\tau) - (a_{F0}+b_{F0}\tau)\Big)^2 \tag{15}$$

ここで，$\mathrm{Period_{F0}}$ (5フレームシフト) は直線近似する期間である．

一方，スペクトル包絡の変形量 $A_s(t)$ 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，次式のように定義する．

$$A_s(t) = \Big(\tfrac{1}{\mathrm{N_{max}}} \sum_{n=1}^{\mathrm{N_{max}}} b_s(n)^2\Big)\Big(\tfrac{1}{\mathrm{N_{max}}} \sum_{n=1}^{\mathrm{N_{max}}} err_s(n)^2\Big) \tag{16}$$

$b_s(n)$ と $err_s(n)$ は，$a_s(n)$ と $b_s(n)$ をパラメータとして次式を最小化することで得られる．

$$err_s(n)^2 = \sum_{\tau=0}^{\mathrm{Period_s}-1} \Big(10\log_{10} Env(n,t-\tau) - (a_s(n)+b_s(n)\tau)\Big)^2 \tag{17}$$

ここで，$\mathrm{Period_s}$ (10フレームシフト) は直線近似する期間である．

3.6 有声休止らしさの評価

有声休止らしさ $P_{fp}(t)$ $(0 \le P_{fp}(t) \le 1)$ は，こうして得た二つの特徴 $A_i(t)$ $(i=f,s)$ の短時間平均

$$S_i(t) = \tfrac{1}{\mathrm{Period_{fp}}} \sum_{\tau=0}^{\mathrm{Period_{fp}}-1} A_i(t-\tau) \quad (i=f,s) \tag{18}$$

に基づいて，

$$P_{fp}(t) = \exp\Big(-\frac{(\mathrm{R}\,S_f(t) + (1-\mathrm{R})\,S_s(t))^2}{\mathrm{W}^2}\Big) \tag{19}$$

のように定義する．ここで，$\mathrm{Period_{fp}}$ (10フレームシフト) は平均する期間であり，R (0.034) は二つの特徴に対する重み付けを決める定数，W (0.575) は主に考慮する変動・変形の範囲を決める定数である．

本手法では，有声休止らしさ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たと判定する．そのために，条件 $P_{fp}(t) > e^{-1}$ を満たし続ける限り $P_{fp}(t)$ を累積加算して，合計値 $Sum_{fp}(t)$ を求める．満たさない場合には $Sum_{fp}(t) = 0$ にリセットする．そして，$Sum_{fp}(t)$ が一定の閾値 $\mathrm{Th_{fp}}(7e^{-1})$ より大きいとき，現在の時刻 t が有声休止区間内であると判定する．

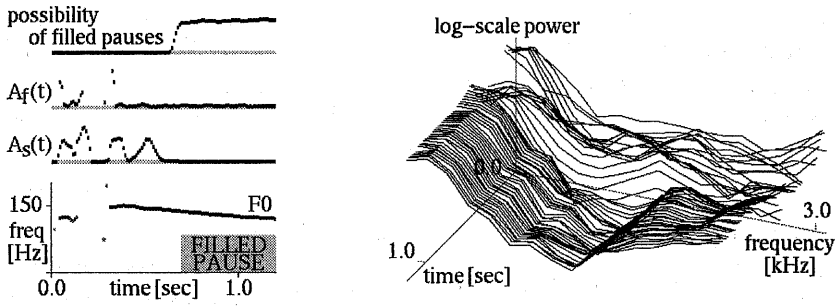

図 3: システムの画面表示例: 男性の自由発話の一部である「一階にー」/iqkaini-/ を入力し終わった時点での，基本周波数等の中間結果と最終出力 (左側) および対応するスペクトル包絡 (右側)

4 システムの実装

音声音響信号を入力し，有声休止らしさとそれに基づく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，提案した手法に基づいて構築した．出力形式として，視覚化のためのコンピュータグラフィックス，聴覚化のための音響信号，音声認識・対話システム等で使用するための連続的に変化する数値 (タイムスタンプ付き) の三種類に対応した．コンピュータグラフィックスの出力では，中間結果や出力が確認できるよう，時間周波数平面上をスクロールする基本周波数と高調波構造の軌跡を表示するウィンドウ，スペクトル包絡を3次元表示するウィンドウ，基本周波数の変動量，スペクトル包絡の変形量，有声休止らしさ，有声休止区間の判定結果を表示するウィンドウなどが表示される．出力音響信号は，推定した高調波構造が適切かどうかを確認できるよう，$Pow(k,t;F_{F0}(t))$ に基づいて，正弦波重畳モデルで合成される．

我々は本システムを分散環境で実装し，音響信号の入出力，3 章で提案した手法の計算，中間結果や出力の視覚化といったシステムを構成する各機能を，LAN (Ethernet) 上に分散した異なるプロセスとして実行できるようにした．その際，システムの拡張や音声認識・対話システム等との接続を容易にするために，RACP (Remote Audio Control Protocol) を設計し，それに基づいて実装した．RACP は，RMCP (Remote Music Control Protocol)[19] を音響信号の伝送用に拡張したネットワークプロトコルである．現在の実装では，提案手法の計算はパーソナルコンピュータ (Pentium MMX 200 MHz CPU, Linux 2.0) 上で実行され，音響信号の入出力や視覚化の処理はワークステーション (SGI Octane R10000 250 MHz CPU, Irix 6.4) 上で実行される．

5 実験結果

日本語の音声対話コーパスから 100 発話を抜粋した評価用音声データを用意し，それを対象に実験をおこなった．使用したコーパスは，Wizard of Oz 方式を用いて収録した自由発声音声の対話コーパス[20] であり，渋谷のレストラン，デパートなどの道案内をタスクとするものである．100 発話は，10 名の話者 (成人男性 5 名，成人女性 5 名) の発話から，有声休止を最低一つ含む発話だけを，各話者につき 10 発話抜粋して構成した．ここでの発話は，無音で区切られた音声区間を意味し，300 ms 以上の無音区間を自動検出して切り出された．

本実験では，再現率 (recall rate) と適合率 (precision rate) の観点から評価をおこなった．以下にこれらの定義を示す．

$$\text{再現率} = \frac{\text{正しく検出した有声休止の数}}{\text{有声休止の総数}} \quad (20)$$

$$\text{適合率} = \frac{\text{正しく検出した有声休止の数}}{\text{有声休止として検出した数}} \quad (21)$$

実験の結果，再現率は 84.9 % (107 / 126)，適合率は 91.5 % (107 / 117) であった．ここでは，システムが有声休止として検出した区間が実際の有声休止区間に含まれているとき，正しく検出したと判定した．正しく検出した例の書き起こしテキストを以下に示す．システムが検出した箇所の後に "@" マークをつけた．

「えー@っと」
「うんとー@」
「はい，では一@」
「あのー@，道を曲がる前にー@」
「えー@と，西武A<e->館を知っていますか」
「あー@んと，パルコパート1<ichi>ごぞんじですか^」
「ルームとしか言ってー@，ないんでお店の名前忘れちゃったんですけど」
「いや，ちょっとわからないんですけどー@，もう一度説明していただけません^」

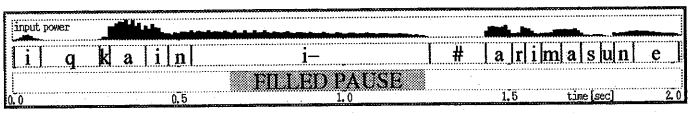

図 4: 有声休止区間の検出例: 男性の自由発話「一階にー，ありますね」/iqkaini-#arimasune/ に対して，人間が手作業でラベリングした音素系列(上段)とシステムが検出した有声休止区間(下段)

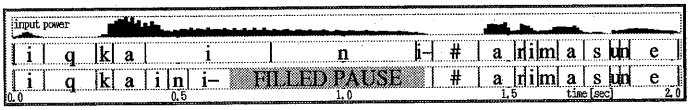

図 5: 音素系列のアラインメントの改善例: 元のアラインメントの結果(上段)と検出した有声休止区間を用いることで改善されたアラインメントの結果(下段)

システムの出力例を図 3に示す．これは，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画面表示である．左側のグラフが，基本周波数(F0) $F_{F0}(t)$，基本周波数の変動量 $A_f(t)$，スペクトル包絡の変形量 $A_s(t)$，有声休止らしさ(possibility of filled pauses) $P_{fp}(t)$，有声休止区間("FILLED PAUSE" と書かれた濃い領域のある区間)を表し，右側のグラフが，対応するスペクトル包絡 $Env(n,t)$ を表す．ここでは，/ni-/の有声休止が適切に検出できている．実際にはこれらの表示はスクロールしており，リアルタイムに確認可能である．次に，音素系列と検出した有声休止区間の対応を調べた例を図 4に示す．図中上段が，男性の自由発話「一階にー，ありますね」/iqkaini-#arimasune/ に対して，人間が音を聞きながら手作業でアラインメントを調節した音素系列を表す．図中下段のシステムが検出した有声休止(音節の引き延ばし)区間が，適切であることがわかる．

検出誤りの原因を分析した結果を以下にまとめる．まず，再現率における誤り(検出もれ, miss)では，有声休止音の持続時間が短すぎたり(短い「えー」/e-/など)，基本周波数の変化が通常より大きすぎたり，声がしわがれて高調波成分が乱れたりしたのが主な原因であった．特に，ある特定の男性話者に関しては，再現率が 53.8％と，例外的に低かった．これは，その話者がほとんどの発話を低いしわがれた声で話すためであった．一方，適合率における誤り(誤検出, false alarm)の多くは，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の箇所で起きていた．そうした有声音は，音韻のなまけ現象(target undershoot)によって生じやすかった．ただし，通常は単語中では基本周波数の変動が十分大きいため，たとえ似た母音が連続することがあっても，そのほとんどは有声休止ではないと正しく判断されていた．

6 音声認識の性能向上へ向けた予備実験

提案した有声休止検出手法を音声認識の枠組みで活用する第一段階として，検出した有声休止区間を用いることで，音素系列のアラインメントが改善されうるかを調査する予備実験をおこなった．本実験では，入力に対応した正解の音素系列は与えるものとし，その系列に合わせて連結した音素HMMと，入力とのマッチングをおこなった．標本化周波数は16 kHz，フレームシフト時間は10 msecとし，音響特徴量としては，16次のLPCメルケプストラム係数とその時間方向の変化量の計32次元を用いた．一般に，有声休止は，ビタービ(Viterbi)アルゴリズムによって決定される音素系列のアラインメントに対して，悪影響を与える．このアラインメントを改善するために，ここでは，状態継続時間長を動的に制御することを試みた．具体的には，検出した有声休止区間において，ある母音の音素から次の音素への状態遷移を抑制した．

こうして音素系列のアラインメントが改善された例を図 5に示す．図中上段が，図 4と同一の発話に対する元のアラインメントの結果を表し，下段が，有声休止区間を用いた場合のアラインメントの結果を表す．実際に，有声休止が含まれる/ni-/の箇所で，アラインメントが改善されていることがわかる．このような結果から，有声休止を含む発話に関しては，有声休止検出手法を用いることで，音声認識システムの性能を向上できる可能性があることがわかる．

7 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，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．本手法では，二つの音響的な特徴量(基本周波数の変動とスペクトル包絡の変形)が共に小さい箇所をボトムアップに検出することで，音韻やつなぎ語の種類を問わずに有声休止を検出することを可能にした．その際，実世界の音響信号中の話者の音声に対しても本手法を適用できるように，入力音響信号中の最も優勢な高調波構造に基づいて，これらの音響的な特徴量を推定した．本手法を実装したシステムを用いて，日本語の音声対話コーパスに対して実験した結果，再現率84.9%，適合率91.5%の高い精度で，リアルタイムに有声休止を検出できることが確認された．

今後は，他の特徴の利用も考慮しながら，有声休止の検出精度を向上させていく予定である．例えば，音楽音響信号を対象としたビートトラッキングの研究[21)~24)]等を応用して話速を捉え，それを特徴として利用した検出の可能性も検討していく．また，音声認識システムで本手法を利用する際に，6章では有声休止区間のみを用いていたが，今後は有声休止らしさを用いることも考慮していく．さらに，対話において大切な役割を持つ有声休止を積極的に活用した音声対話システムの構築もおこなっていく予定である．

謝辞

本研究は，通商産業省RWCプロジェクトの一環として電子技術総合研究所RWIセンターで実施した．同プロジェクト，同センターで研究推進に当たられた方々に感謝する．

参考文献

[1] 田窪行則: 音声言語の言語学的モデルをめざして — 音声対話管理標識を中心に —, 情報処理学会誌, Vol. 36, No. 11, pp. 1020–1026 (1995).

[2] 伊藤克亘: 音声対話システム, 自然言語処理 — 基礎と応用 —, 電子情報通信学会, pp. 302–322 (1999).

[3] Ward, W.: Understanding Spontaneous Speech: The Phoenix System, *Proc. of ICASSP 91*, pp. 365–367 (1991).

[4] 中川聖一, 小林聡: 自然な音声対話における間投詞・ポーズ・言い直しの出現パターンと音響的性質, 日本音響学会誌, Vol. 51, No. 3, pp. 202–210 (1995).

[5] Kai, A. and Nakagawa, S.: Investigation on unknown word processing and strategies for spontaneous speech understanding, *Proc. of Eurospeech '95*, pp. 2095–2098 (1995).

[6] 甲斐充彦, 中川聖一: 冗長語・言い直し等を含む発話のための未知語処理を用いた音声認識システムの比較評価, 電子情報通信学会論文誌 D-II, Vol. J80-D-II, No. 10, pp. 2615–2625 (1997).

[7] O'Shaughnessy, D.: Recognition of hesitations in spontaneous speech, *Proc. of ICASSP 92*, pp. I–521–524 (1992).

[8] Quimbo, F. C. M., Kawahara, T. and Doshita, S.: Prosodic analysis of fillers and self-repair in Japanese speech, *Proc. of ICSLP 98* (1998).

[9] 田中敏: 「休止」の意味論, 言語, Vol. 22, No. 8, pp. 20–27 (1993).

[10] Rose, R. L.: *The communicative value of filled pauses in spontaneous speech*, PhD Thesis, University of Birmingham (1998).

[11] 伊藤克亘, 秋葉友良, 上條俊一, 田中和世: 休止を区切りとした対話処理, 情処研報音声言語情報処理 95-SLP-7-21, pp. 135–138 (1995).

[12] 峯松信明, 片岡嘉孝, 中川聖一: 講演調の話し言葉に対する分析, 情処研報音声言語情報処理 95-SLP-8-7, pp. 39–46 (1995).

[13] Flanagan, J. L. and Golden, R. M.: Phase Vocoder, *The Bell System Technical J.*, Vol. 45, pp. 1493–1509 (1966).

[14] Boashash, B.: Estimating and Interpreting The Instantaneous Frequency of a Signal, *Proc. of the IEEE*, Vol. 80, No. 4, pp. 520–568 (1992).

[15] Charpentier, F. J.: Pitch detection using the short-term phase spectrum, *Proc. of ICASSP 86*, pp. 113–116 (1986).

[16] 阿部敏彦, 小林隆夫, 今井聖: 瞬時周波数に基づく雑音環境下でのピッチ推定, 電子情報通信学会論文誌 D-II, Vol. J79-D-II, No. 11, pp. 1771–1781 (1996).

[17] 河原英紀, 片寄晴弘, Patterson, R. D., de Cheveigné, A.: 瞬時周波数を用いた基本周波数の高精度の抽出について, 日本音響学会音楽音響研究会 H-98-116, pp. 31–38 (1998).

[18] Abe, T., Kobayashi, T. and Imai, S.: The IF spectrogram: a new spectral representation, *Proc. of ASVA 97*, pp. 423–430 (1997).

[19] 後藤真孝, 根山亮, 村岡洋一: RMCP: 遠隔音楽制御用プロトコルを中心とした音楽情報処理, 情報処理学会論文誌, Vol. 40, No. 3, pp. 1335–1345 (1999).

[20] Itou, K., Akiba, T., Hasegawa, O., Hayamizu, S. and Tanaka, K.: A Japanese spontaneous speech corpus collected using automatically inferencing Wizard of OZ system, *J. Acoust. Soc. Jpn. (E)*, Vol. 20, No. 3 (1999).

[21] 後藤真孝, 村岡洋一: ビートトラッキングシステムの並列計算機への実装 — AP1000によるリアルタイム音楽情報処理 —, 情報処理学会論文誌, Vol. 37, No. 7, pp. 1460–1468 (1996).

[22] 後藤真孝, 村岡洋一: 音響信号を対象としたリアルタイムビートトラッキングシステム— コード変化検出による打楽器音を含まない音楽への対応 —, 電子情報通信学会論文誌 D-II, Vol. J81-D-II, No. 2, pp. 227–237 (1998).

[23] Goto, M. and Muraoka, Y.: Real-time Beat Tracking for Drumless Audio Signals: Chord Change Detection for Musical Decisions, *Speech Communication*, Vol. 27, No. 3–4, pp. 311–335 (1999).

[24] 後藤真孝: 音楽音響信号を対象としたリアルタイムビートトラッキングに関する研究, 博士論文, 早稲田大学 理工学部 (1998).