自然発話中の有声休止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[†]

A Real-Time System Detecting Filled Pauses in Spontaneous Speech

Masataka GOTO[†], Katunobu ITOU[†], and Satoru HAYAMIZU[†]

あらまし　本論文では，代表的ないいよど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，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有声休止に対処する研究事例はあったが，いいよど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に着目し，音韻的に変化が少ない持続した有声音（有声休止音）を検出する．その際，ボトムアップな信号処理によって，有声休止音がもつ二つの音響的特徴（基本周波数の変動が小さくスペクトル包絡の変形が小さい）を検出することで，トップダウン情報を使わない言語非依存な検出を可能とする．本手法をリアルタイムに実行するシステムを実装し，有声休止箇所のマーク付け作業を施した日本語の音声対話コーパスを用意して，評価実験を行った．その結果，30 名の話者の自然発話に対し，F 値 0.726 の精度で有声休止を検出できることが確認された．

キーワード　有声休止，いいよどみ，つなぎ語，自然発話，対話

1. ま え が き

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，話し言葉特有のいいよど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 1 段階として，本論文では，有声休止（filled pause）と音節の引き延ばし（word lengthening）の二つのいいよど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，いいよど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，いいよど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，いいよどみ現象を含まない，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，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得るため，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(注 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，いいよど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．

[†] 電子技術総合研究所，つくば市
Electrotechnical Laboratory, Tsukuba-shi, 305–8568 Japan

(注 1)：多くの論文では，いいよどむ際に用いる「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，本論文ではつなぎ語（filler）という用語を用いる．

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，人間はなじみのない外国語に対しても，韻律的な手がかりからいいよど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とどまっており，自動的に有声休止を検出するシステムは構築されていなかった．

本論文では，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いいよどみ現象を検出する手法を提案する．両者のいいよどみ現象は同様な音響的特徴をもっており，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，本論文では以下,「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 **2.** で有声休止の役割を考察し，**3.** で提案手法のアルゴリズムを説明する．次に，**4.** で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，**5.** で評価実験の結果を示す．最後に，**6.** でまとめを述べる．

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現れる現象であるからだと考える．話者が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もつ．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行うことで，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．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かかわらず発話権をもち続けたいとき（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行うことで，聴取者に次の発話を待ってほ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（ちゅうちょ），謙遜（そん）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．更に，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．その際には，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行うこともある．

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 1 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行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，スロットに入り得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，前述の発話権の保持等の機能を効果的に使う目的で，有声休止を導入していくとよい．

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.** 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，話者が調音器官（喉頭を含む）の位置・状態を有声休止中に変化させることは難しい．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．

以上から本手法では，有声休止音がも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

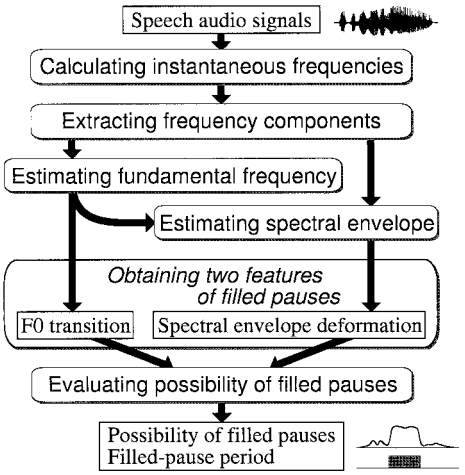

図 1 有声休止検出手法の処理の流れ
Fig. 1 Overview of our filled-pause-detection method.

機能するように，LPC 等の単一音源を前提とした分析は行わず，最も優勢な高調波構造に基づく推定を行う．そして，前述した有声休止音の二つの特徴を定量的にとらえ，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

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

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

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．

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

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 ms 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]．

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

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．

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

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したい．そこで，非周期的な雑音に加え，高調波構造をも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，本論文では以下，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_{Hz} を，次のように cent で表された周波数 f_{cent} に変換する．

$$f_{\mathrm{cent}} = 1200\log_2 \frac{f_{\mathrm{Hz}}}{440 \times 2^{\frac{3}{12}-5}} \tag{6}$$

基本周波数の可能性 $P_{F0}(F,t)$ は，次式のように定義する．

$$P_{F0}(F,t) = \int_{-\infty}^{\infty} p(x;F)\ \Psi'_p(x,t)\ dx \tag{7}$$

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）．

$$p(x;F) = \sum_{h=1}^{\mathrm{N}} c(h)\ G(x;F + 1200\log_2 h, \mathrm{W_f}) \tag{8}$$

$$G(x;m,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \tag{9}$$

ここで，N は考慮する高調波成分の数（基本周波数成分も数える），$\mathrm{W_f}$ 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 $c(h) = \mathrm{H}^{h(h-1)}$（H は定数）とする．現在の実装では，N は，音の高さの知覚で主に用いられる周波数帯域 [19] を十分にとらえられるように 8 とし，$\mathrm{W_f}$ は，周波数成分の高調波関係のずれと分離知覚との関係 [20] を考慮して 17 cent とした．また，H は，優勢の判断に対する低次の高調波成分の寄与がやや大きくなるように，予備実験により 0.9849 とした．

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．

$$F_{F0}(t) = \operatorname*{argmax}_F\ P_{F0}(F,t) \tag{10}$$

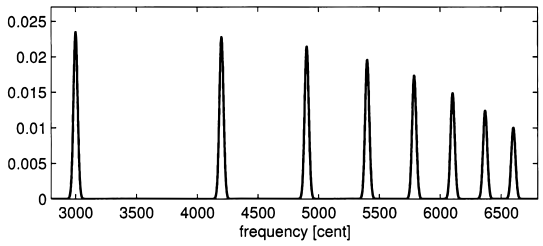

図 2　基本周波数 F の高調波成分だけを通過させるフィルタ関数 $p(x;F)$（$F = 3000$ cent の場合を図示）

Fig. 2　The filter function $p(x;F)$ passing harmonic components of the F0 F. This example shows the frequency response when $F = 3000$ cent.

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．

$$Pow(k,t;F_{F0}(t)) = \max_x G(x;F_{F0}(t)+1200\log_2 k,\mathrm{W_s})\ \Psi'_p(x,t) \tag{11}$$

ここで，$\mathrm{W_s}$ はガウス分布の標準偏差を表す．$\mathrm{W_s}$ は，周波数成分の高調波関係のずれと分離知覚との関係 [20] を考慮しながら，基本周波数の推定結果に多少の誤差があっても適切にパワーを得られるように，$\mathrm{W_f}$ よりも大きい 35 cent とした．

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 1，第 2 フォルマントをとらえられるような上限周波数（3200 Hz）を設けて行う．有声休止音の特徴としては，包絡の大局的な変形をとらえた方が良いため，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（$1 \leqq n \leqq \mathrm{N_{max}}$ (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

$$\sum_{n=1}^{\mathrm{N_{max}}} Env(n,t) = 1 \tag{12}$$

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 2 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．

$$A_f(t) = \left| b_{F0} \right| \tag{13}$$

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．

$${err_{F0}}^2 = \sum_{\tau=0}^{\mathrm{Period_{F0}}-1} \left(F_{F0}(t-\tau) - (a_{F0}+b_{F0}\tau)\right)^2 \tag{14}$$

ここで，$\mathrm{Period_{F0}}$ は直線近似する期間であり，一つの音韻区間内の局所的な傾きをとらえられるように，日本語母音の平均継続時間（64〜101 ms）[21] よりも十分短い 5 フレームシフト（50 ms）とした．

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 2 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．

$$A_s(t) = \left(\frac{1}{\mathrm{N_{max}}}\sum_{n=1}^{\mathrm{N_{max}}} b_s(n)^2\right) \cdot \left(\frac{1}{\mathrm{N_{max}}}\sum_{n=1}^{\mathrm{N_{max}}} err_s(n)^2\right) \tag{15}$$

$b_s(n)$ 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．

$$err_s(n)^2 = \sum_{\tau=0}^{\mathrm{Period_s}-1} \left(10\log_{10} Env(n,t-\tau) - (a_s(n)+b_s(n)\tau)\right)^2 \tag{16}$$

ここで，$\mathrm{Period_s}$ は直線近似する期間であり，音韻間の遷移による包絡の変形が適切に反映されるよう 10 フレームシフト（100 ms）とした．

3.6 有声休止らしさの評価

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

$$S_i(t) = \frac{1}{\mathrm{Period_{fp}}}\sum_{\tau=0}^{\mathrm{Period_{fp}}-1} A_i(t-\tau) \tag{17}$$

に基づいて，

$$P_{fp}(t) = \exp\left(-\frac{(\mathrm{R}\ S_f(t)\ +\ (1-\mathrm{R})\ S_s(t))^2}{\mathrm{W^2}}\right) \tag{18}$$

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る．R(0.011) は二つの特徴

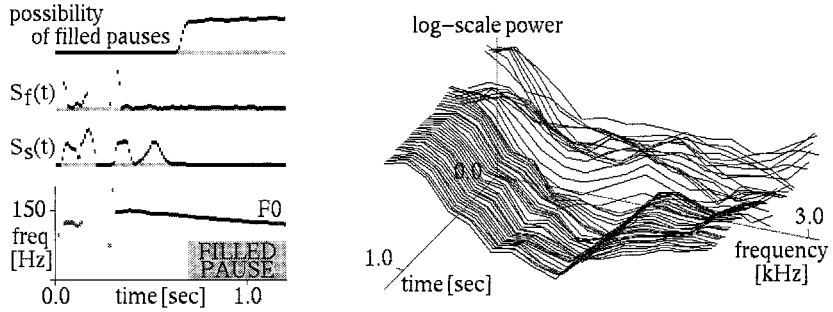

図 3　システムの画面表示例：男性の自由発話の一部である「一階にー」/iqkaini-/ を入力し終わった時点での，基本周波数等の中間結果と最終出力（左側）及び対応するスペクトル包絡（右側）

Fig. 3　An example of the F0 and intermediate results (left) and the corresponding spectral envelope (right) for part of a male spontaneous utterance /iqkaini-/.

に対する重み付けを決める定数，W(0.39) は主に考慮する変動・変形の範囲を決める定数であり，実験的に値を定めた．

本手法では，有声休止らしさが十分高い値をとり続けるときに，話者が有声休止を行っていると判定する．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，累積値 $Sum_{fp}(t)$ 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，$Sum_{fp}(t)$ が一定のしきい値 $\mathrm{Th_{fp}}(9.5e^{-1})$ 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．なお，$\mathrm{Th_{fp}}$ の設定を大きくすることで，検出もれ（miss）を増やす代わりに誤検出（false alarm）を減らしたり，小さくすることで，誤検出を増やす代わりに検出もれを減らしたりすることが可能である．このトレードオフについては，**5. 5** で考察する．

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 3 種類に対応した．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．

コンピュータグラフィックスの出力の画面表示例を図 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 $S_f(t)$，スペクトル包絡の変形量 $S_s(t)$，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．

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.** 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）[22] 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

システムの有効性を確認するために，音声対話コーパスを用いて評価実験を行った．有声休止検出システムの評価の枠組みは確立していなく，多くの音声対話コーパスでは，有声休止箇所が明示的にマーク付けさ

れていることは少ないため，直接評価に用いることはできない．そこで本研究では，既存の音声対話コーパスに対するマーク付けの方法について検討し，それらを書き起こしに加える作業を実施した．そして，マーク付けした箇所とシステムの出力を比較することで，評価を行った．

以下，**5.1** で具体的なマーク付け作業について説明し，**5.2** で有声休止の出現傾向や継続時間を分析する．そして，**5.3** でシステムの評価基準を示し，**5.4** で実際にシステムを評価した結果を述べる．最後に，**5.5** でしきい値 Th_{fp} の変更によるシステム性能の変化に関して考察する．

5.1 音声対話コーパスに対するマーク付け

本評価では，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，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 [23] を用いた．被験者 30 名による 150 対話（被験者 1 人につき 5 種類のタスク），全 4697 発話（音声区間の自動切出しに成功した発話のみ使用）を対象とした．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．発話長は，平均 1.16 s，標準偏差 0.89 s，最小 0.10 s，最大 8.62 s であった．

本コーパスを用いて評価するためには，有声休止箇所をマーク付け（記号付与）しておく必要がある．そこで，表 1 の記号を定義し，音声を聞きながら書き起こしテキスト中の該当箇所の直後（記号 [ ] は前後）に挿入する作業を，2 名が相互にチェックし合う形式で実施した．しかし，有声休止の有無はそもそも 2 値で判断できるものではなく，判断が微妙で迷うような状況も多い連続的な現象である．そこで，そうした判断が微妙な箇所には，記号 ： _ を付けた．「いいよどみ以外」とは，主に口癖による文節末の音節の引き延ばしなどである．記号 [ ] は，有声休止の有無にかかわらず，言語的な役割がつなぎ語である箇所を囲むように付けた．記号を付与した書き起こしテキストの例を以下に示す．

「それではいいです．[えー@っとー@]」
「[えー@] 東急ハンズの一階にありま~す」
「の場所を：知りたいん@，ですけど」
「つぼ八@，へ行きたいんですけど_，どこでしょうか」

更に，音響信号中で，記号 @ ： ~ _ のそれぞれに該当する区間にマークを付けるための専用エディタを開発し，エディタ上で区間の開始時刻と終了時刻を指定する作業を行った．エディタの画面表示例を図 4 に示す．図中上段は音響パワーの時間変化で，下段は書き起こしテキストを用いて音素系列を仮にアラインメントした結果である．図中中段が，これらを参考にしながら実際に作業を行うための領域で，指定した開始時刻と終了時刻の間の区間が，記号 @ ： ~ _ ごとに異なる色で表示される．部分的に音響信号を再生しながら作業でき，発話単位でのカーソル移動にも対応するなど，効率の良い作業環境を提供している．

表 1 有声休止箇所を示す記号
Table 1 Symbols for indicating filled pauses.

現象			記号
有声休止（音節の引き延ばし）	いいよどみ	明らか	@
		微妙	:
	いいよどみ以外	明らか	~
		微妙	_
つなぎ語（filler）			[ ]

5.2 有声休止の出現形態の分析

記号を付与した書き起こしテキストに基づき，5 種

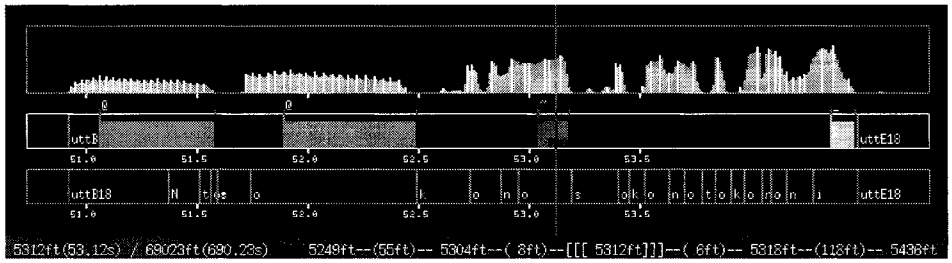

図 4 有声休止箇所の記号の区間指定用エディタの画面表示例
Fig. 4 Screen snapshots of our editor program for marking the duration of filled-pause symbols.

類のタスクの各々について，各記号を含む発話数の全発話数に対する割合を分析した結果を図 5 に示す．これから，タスク 5 では，明らかないいよどみ（@）を含む発話の割合が多いことがわかる．これは，タスク 1〜4 が被験者の質問にシステムが答えて道案内をするのに対し，タスク 5 は被験者の側が道案内をするという違いがあるためである．そのため，想起・思考しながら発話することが多くなり，いいよどみが増えたと考えられる．なお，つなぎ語内とそれ以外を別々に分析した結果も，同様の傾向を示した．つまり，タスク 5 では，つなぎ語内の有声休止だけが増えるのではなく，発話中の有声休止全体が増えていたことになる．

次に，エディタでマーク付けした区間に基づき，各記号の継続時間を分析した結果を図 6 に示す．これから，いいよどみの区間（@ :）の方がいいよどみ以外の区間（~ _）よりも長く,「明らか」と判断された区間（@ ~）の方が「微妙」と判断された区間（: _）よりも長い傾向にあることがわかる．ただし，記号間で分布の重なりが大きいため，継続時間は現象の識別基準としては十分でない．

5.3 評価方法

マーク付けした区間とシステムが出力した有声休止区間を比較して評価する．評価基準を以下に示す．

(i)　システムの出力区間が，記号 @ の区間と重なりをもつ場合，正解（正しく検出した）と判定する．

(ii)　記号 : の区間に関しては，人間の判断も揺れるような微妙な箇所なので，システムが検出した場合には正解に含めるが，検出できなくても検出もれ（miss）とはみなさない．つまり，システムが検出した場合だけ評価の考慮に入れる．

(iii)　記号 ~ _ の区間に関しては，システムが検出してもしなくても，本評価では考慮しない．口癖等による音節の引き延ばしは，いいよどみによる有声休止とは言語的に現れる品詞が異なる傾向にあるという報告 [24] を考慮し，今後，言語情報と併用して識別していく必要があるためである．

これらに基づき，再現率（recall rate），適合率（precision rate），及び両者を統合した F 値（F-measure）[25] の観点から評価を行った．以下に定義を示す．ここで，有声休止の総数とは，評価基準 (i) の記号 @ の区間の数と，評価基準 (ii) のシステムが検出した記号 : の区間の数の合計である．F 値では，再現率と適合率を等価に扱いたいため $\beta = 1$ とした．

$$\text{再現率 } (R) = \frac{\text{正しく検出した有声休止の数}}{\text{有声休止の総数}} \quad (19)$$

$$\text{適合率 } (P) = \frac{\text{正しく検出した有声休止の数}}{\text{有声休止として検出した数}} \quad (20)$$

$$\text{F 値} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad (21)$$

なお，本評価で用いた音声対話コーパスとマーク付けした区間は，**3.6** の定数の設定では用いられたが，**3.1** から **3.5** までの手法の導出及び定数の設定には用いられていない．

5.4 評価結果

まず，話者ごとの F 値の評価結果と，比較のための平均モーラ長（発話時間と対応するローマ字の書き起こしから推定）との関係を図 7 に示す．平均モーラ長を示したのは，話速が検出精度に影響する可能性があるからである．この図から，話者ごとの検出精度の違いが大きいことがわかる．そこで，F 値が悪い 1，26，30 番の話者に関して主な理由を調査した．話者 1 では，話速がやや早く有声休止の長さが短い傾向にあったため，検出もれが多く再現率が低かった．誤検出（false alarm）は少ないが，正しく検出できた数も少ないために結果として適合率と F 値が低くなった．話者 26 に関しては，そもそも有声休止箇所が少ないことに加え，数箇所の有声休止をややしわがれ気味の声や小さい声で発声していたために検出できず，再現率が低かった．

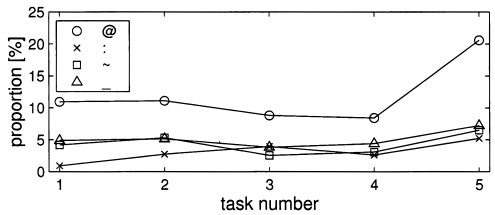

図 5　有声休止を含む発話の割合
Fig. 5　Proportion of the number of utterances containing filled pauses.

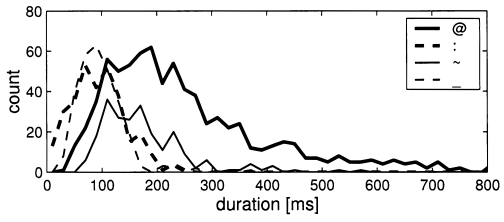

図 6　有声休止の継続時間のヒストグラム
Fig. 6　Histogram of the duration of filled pauses.

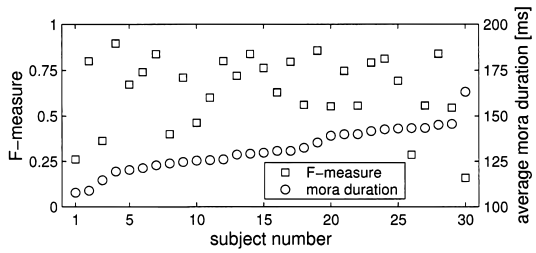

図 7　話者ごとの F 値と平均モーラ長の関係
Fig. 7　Relationship between F-measure and average mora duration.

表 2　全話者を通じた評価結果
Table 2　Results of evaluating the system on all the subjects.

	再現率	適合率	F 値
条件 1（30 名の話者）	0.754	0.700	0.726
条件 2（28 名の話者）	0.764	0.741	0.752

表 3　基本周波数の変動量とスペクトル包絡の変形量のいずれか一方だけを用いた場合の評価結果
Table 3　Results of evaluating the system that uses either fundamental frequency transition or spectral envelope deformation.

用いた特徴	再現率	適合率	F 値
基本周波数の変動量のみ	0.427	0.364	0.393
スペクトル包絡の変形量のみ	0.789	0.639	0.706

また，フラットな基本周波数（F0）の「いいえ」が数箇所あるために適合率も低く，F 値が低くなった．話者 30 は話速が非常に遅く，音節が引き延ばされる傾向にあったために，誤検出が多かった．そのため適合率が 0.095 と極めて低く，F 値も低くなった．

次に，全話者を通じた評価結果を表 2 に示す．各話者の総発話時間が異なるため，総発話時間で重み付けして正規化し，評価値を求めた．条件 1 は，全話者を用いた場合である．条件 2 は，これから最も平均モーラ長が短い話者 1 と最も平均モーラ長が長い話者 30 を除いた場合で，評価結果が向上していることがわかる．これは，話速を自動検出して考慮していくことで，精度を向上できる可能性を示唆している．

本手法では，基本周波数の変動量とスペクトル包絡の変形量に基づいて有声休止を検出しているが，これら二つの特徴量に対する重み付けを決める定数 R を 0 あるいは 1 に設定することで，いずれか一方だけを用いた場合の評価を行うことができる．その結果を表 3 に示す．この結果から，スペクトル包絡の変形量のみを用いた場合（$R = 0.0$, $W = 0.33$ の場合）の方が，基本周波数の変動量のみを用いた場合（$R = 1.0$, $W = 4.0$ の場合）よりも評価結果の値が大幅に高いことがわかる．これは，基本周波数の推定結果に基づいてスペクトル包絡を求めているため，スペクトル包絡の変形量が基本周波数の変動量と完全には独立でない，つまり，基本周波数の変動の影響でスペクトル包絡の変形量が大きくなることがあり得るからである．そこで結果的に，スペクトル包絡の変形量のみを用いても基本周波数の変動をある程度考慮した検出がなされ，全話者を平均すると，F 値が表 2 の条件 1 に比較的近い値となっている．ただし，話者によっては，基本周波数の変動量を用いないと誤検出が大幅に増えて F 値が 0.1 以上も下がるため(注 2)，両者の特徴量をともに用いて検出することが有効である．

本システムの主な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）では，有声休止音の持続時間が短すぎたり（短い「えー」など），声がしわがれて高調波成分が乱れたり，パワーが弱く無声音に近い状態でつなぎ語を発声したりしたのが主な原因であった．また，基本周波数の変化が通常より大きすぎて誤ることもあり，他の感情・ニュアンスが混ざりながら文節末で有声休止する箇所で起きやすかった．一方，適合率における誤り（誤検出）の多くは，平たん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（例えば「**いい**え」「どう**いう**」「食べるもの**のお**店」など）で起きていた．そうした有声音は，音韻のなまけ現象（target undershoot）によって生じやすく，特定の単語や話者に誤りが集中する傾向があった．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，その多くは有声休止ではないと正しく判断されていた．ほかにも，話速が局所的に遅くなる箇所では各音韻の持続時間が長くなり，誤検出が起きやすかった．こうした例外的な発声にも対処していくには，より多くの音響的な特徴や言語情報を統合していく必要がある．

5.5　再現率と適合率のトレードオフ

有声休止らしさの累積値 $Sum_{fp}(t)$ に対するしきい値 $\mathrm{Th_{fp}}$ は，再現率（recall rate）と適合率（precision

(注 2)：スペクトル包絡の変形量のみを用いた場合には，似た母音が連続する箇所や単語中の長母音の箇所を誤検出しやすくなる．通常，こうした箇所は基本周波数が変動するため，基本周波数の変動量をともに用いることで誤検出が防げる．

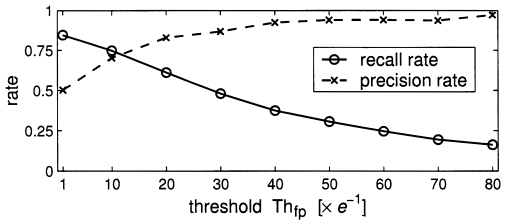

図 8　しきい値 $\mathrm{Th_{fp}}$ を変化させたときの評価結果
Fig. 8　Results of evaluating the system under different thresholds $\mathrm{Th_{fp}}$.

rate）のトレードオフを調整するための重要な定数である．そこで，この値を変化させたときの再現率と適合率の評価結果を図 8 に示す．これから，適合率を高くしたい場合には $\mathrm{Th_{fp}}$ に大きな値を設定し，再現率を高くしたい場合には $\mathrm{Th_{fp}}$ に小さな値を設定すればよいことがわかる．

適合率と再現率のどちらをより重視するかは，有声休止の検出結果をどのような用途に用いるかによって異なるため，このようにトレードオフを調整できることは望ましい．例えば，音声認識システムの性能向上のために有声休止検出結果を利用する場合には，適合率が高いことが求められる．その場合，$\mathrm{Th_{fp}}$ を $40e^{-1}$ に設定すれば適合率 0.926 と高い性能が得られるので，音声認識システムの性能向上のために十分使用可能である．このときの再現率は 0.375 となってしまうが，音声認識に悪影響を与えることが予想されるような，400 ms 以上のかなり長い継続時間をもつ有声休止のみを対象に再現率を求めると 0.911 となり，性能向上に寄与することが期待できる．

6. むすび

本論文では，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）がとも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，30 名の話者の自然発話に対し，F 値 0.726 の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，話速等の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ことを予定している．更に，有声休止検出結果を利用した音声認識システムの構築や，対話において大切な役割をも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行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あたられた方々に感謝する．

文　献

[1] 田窪行則，“音声言語の言語学的モデルを目指して — 音声対話管理標識を中心に，” 情報処理，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，pp.302–322, 電子情報通信学会，1999.
[3] W. Ward, “Understanding spontaneous speech: The Phoenix system,” Proc. 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，” 音響誌, vol.51, no.3, pp.202–210, 1995.
[5] A. Kai and S. Nakagawa, “Investigation on unknown word processing and strategies for spontaneous speech understanding,” Proc. 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，” 信学論（D-II），vol.J80-D-II, no.10, pp.2615–2625, Oct. 1997.
[7] D. O’Shaughnessy, “Recognition of hesitations in spontaneous speech,” Proc. ICASSP 92, pp.I-521–524, 1992.
[8] F.C.M. Quimbo, T. Kawahara, and S. Doshita, “Prosodic analysis of fillers and self-repair in Japanese speech,” Proc. ICSLP 98, vol.7, pp.3313–3316, 1998.
[9] 田中　敏，“「休止」の意味論，” 言語，vol.22, no.8, pp.20–27, 1993.
[10] R.L. Rose, The communicative value of filled pauses in spontaneous speech, Master’s 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，” 情処学音声言語情報処理研報，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，” 情処学音声言語情報処理研報，95-SLP-8-7, pp.39–46, 1995.
[13] J.L. Flanagan and R.M. Golden, “Phase vocoder,” The Bell System Technical J., vol.45, pp.1493–1509,

1966.
[14] B. Boashash, “Estimating and interpreting the instantaneous frequency of a signal,” Proc. IEEE, vol.80, no.4, pp.520–568, 1992.
[15] F.J. Charpentier, “Pitch detection using the short-term phase spectrum,” Proc. 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,” 信学論（D-II），vol.J79-D-II, no.11, pp.1771–1781, Nov. 1996.
[17] 河原英紀，片寄晴弘，R.D. Patterson, and A. de Cheveigné, “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,” 音響学音楽音響研資，H-98-116, pp.31–38, 1998.
[18] T. Abe, T. Kobayashi, and S. Imai, “The IF spectrogram: A new spectral representation,” Proc. ASVA 97, pp.423–430, 1997.
[19] R.J. Ritsma, “Frequencies dominant in the perception of the pitch of complex sounds,” J. Acoust. Soc. Am., vol.42, no.1, pp.191–198, 1967.
[20] 柏野邦夫，田中英彦，“二つの周波数成分の分離知覚に関する工学的モデル—複数の要因の評価と統合,” 信学論（A），vol.J77-A, no.5, pp.731–740, May 1994.
[21] 村上仁一，嵯峨山茂樹，“自由発話音声における音響的な特徴の検討,” 信学論（D-II），vol.J78-D-II, no.12, pp.1741–1749, Dec. 1995.
[22] 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,” 情処学論，vol.40, no.3, pp.1335–1345, 1999.
[23] K. Itou, T. Akiba, O. Hasegawa, S. Hayamizu, and K. Tanaka, “A Japanese spontaneous speech corpus collected using automatically inferencing Wizard of OZ system,” J. Acoust. Soc. Jpn. (E), vol.20, no.3, 1999.
[24] 千田恭子，伊藤克亘，後藤真孝，“道案内 WOZ システムとの対話における言い淀み表現の分析,” 言語処理学会第 6 回年次大会，pp.344–347, March 2000.
[25] C.J. van Rijsbergen, Information Retrieval, 2nd ed., Butterworths, 1979.

（平成 12 年 2 月 22 日受付，6 月 30 日再受付）

後藤　真孝　（正員）

1993 早大・理工・電子通信卒．1998 同大大学院博士後期課程了．同年，電子技術総合研究所に入所し，現在に至る．博士（工学）．音楽情報処理，音声言語情報処理，マルチモーダルインタラクションなどに興味をもつ．1992 jus 設立 10 周年記念 UNIX 国際シンポジウム論文賞，1993 NICOGRAPH’93 CG 教育シンポジウム最優秀賞，1997 情報処理学会山下記念研究賞，1999 平成 10 年電気関係学会関西支部連合大会奨励賞各受賞．情報処理学会，日本音響学会，日本音楽知覚認知学会，ICMA，ISCA 各会員．

伊藤　克亘

1988 東工大・工・情報卒．1993 同大大学院博士課程了．現在，電総研知能情報部．音声対話の研究に従事．情報処理学会，日本音響学会会員．工博．

速水　　悟　（正員）

1978 東大・工・産業機械卒．1981 同大大学院修士課程機械工学専攻了．同年，電子技術総合研究所入所．1989〜1990 カーネギーメロン大学，1994 LIMSI/CNRS 客員研究員．工学博士．音声認識，音声対話，人工物とのコミュニケーションに関する研究に従事．日本音響学会，言語処理学会，日本機械学会，IEEE，ISCA 各会員．