音声スポッタ：人間同士の会話中に音声認識が利用可能な音声入力インタフェース

後藤真孝[†] 北山広治[††,☆]
伊藤克亘[†††,☆☆] 小林哲則[††]

本論文では，人間同士の会話中に音声認識システムへ音声コマンドを入力できる「音声スポッタ」という音声インタフェース機能を提案する．従来，会話中のユーザの音声が，音声認識システムと会話相手の人のどちらに対する発話かを，マイク入力による音声だけから識別することは困難だったため，人間同士の会話中に音声認識システムは利用されていなかった．音声スポッタでは，音声に含まれる非言語情報の中から，有声休止（「えー」のように母音の引き延ばし）による言い淀みと，声の高さの2種類を活用することで，各発話が音声認識システムに入力されるかどうかを，ユーザが意図的に制御できるようにする．具体的には，母音を延ばして言い淀んだ後に故意に高い声で発声された特殊な（不自然な）発話だけを音声認識対象と見なし，通常の会話中の発話は無視することで会話の支援を実現する．その応用例として我々は，会話中のユーザに各種情報支援をする「オンデマンド会話支援システム」と，電話での通話中にユーザがBGMを選曲・再生できる「BGM付き電話システム」の2つを構築した．音声スポッタによる発話の検出性能の評価結果やこれらのシステムの試用を通じて，本機能が頑健で便利であることを確認した．

Speech Spotter: Speech Input Interface Capable of Using Speech Recognition in the Midst of Human-Human Conversation

Masataka Goto,[†] Koji Kitayama,[††,☆] Katunobu Itou[†††,☆☆] and Tetsunori Kobayashi[††]

This paper describes a speech-interface function, called *"Speech Spotter"*, which enables a user to enter voice commands into a speech recognizer in the midst of natural human-human conversation. In the past, it has been difficult to use automatic speech recognition in human-human conversation since it was not easy to judge, from only microphone input, whether a user was speaking to another person or a speech recognizer. We enable a user to *intentionally* control whether each utterance is to be accepted (processed) by the speech recognizer by using two kinds of nonverbal speech information: a filled pause (a vowel-lengthening hesitation like "er...") and voice pitch. Speech Spotter regards a user utterance as a command utterance only when it is uttered with a high pitch just after a filled pause. In other words, this function accepts this specially-designed unnatural utterance only and ignores other normal utterances. By using Speech Spotter, we have built two application systems: an on-demand information system for assisting human-human conversation and a music-playback system for enriching telephone conversation. The results from evaluating this function and using these systems have shown that Speech Spotter is robust and convenient enough to be used in face-to-face or cellular-phone conversations.

† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
†† 早稲田大学
Waseda University
††† 名古屋大学
Nagoya University
☆ 現在，株式会社東芝
Presently with Toshiba Corporation
☆☆ 現在，法政大学
Presently with Hosei University

1. はじめに

本論文では，人間同士が日常会話をしている最中に，音声認識によって計算機による支援を可能にすることを目的として，マイク入力だけで人間同士の会話中の音声認識対象区間を指定する手法を検討する．人間同士の会話中に，あたかもそこに第三者がいるかのように計算機の支援を受けられると便利である．たとえば，人と会話をしながら今日が何日かを知りたくなったり，

明日の天気予報やスポーツの結果を知りたくなったりしたときに，通常は会話を中断して近くにいる第三者に尋ねるか，計算機に向かって調査をして解決することが多い．その代わりに，もし計算機が人間同士の会話をモニタリングしていて，知りたいタイミングで情報を教えてくれると，会話を中断することなく各種情報支援が得られて有用である．

しかし，従来の技術では，マイク入力でこのような会話中の情報支援を実現することは難しかった．音声情報のみで実現する場合，ワードスポッティング技術でキーワード検出して支援する方法が考えられる[1)~4)]．しかし，検出されたキーワードが，計算機に対する「処理対象音声」だったのか，会話相手に対する「通常会話音声」だったのかを，事前に話題を限定せずに識別することはできなかった．一方，音声以外の情報を併用する方法では，ボタンを押す（マイクのスイッチを入れる）ことによって識別する方法[5)]や，画像情報からユーザの顔や視線の方向を認識して識別する方法[6),7)]が提案されている．さらに，画像情報を利用しつつ，会話内容を理解して情報支援を行う方法[7),8)]も提案されている．しかし，これらの方法はボタンやカメラといった他の入力デバイスの併用が前提であり，本研究で目指すような，マイク入力だけで識別することはできなかった．

そこで我々は，母音を延ばして言い淀んだ後に故意に高い声で発声する特殊な発話だけを，音声認識対象と見なす新たな音声インタフェース機能「音声スポッタ」を提案する．母音の引き延ばし（有声休止）の直後に高い声で発声するのは不自然な発話であり，人間同士の会話中に現れない．そのため，音声スポッタでは，人間同士の「通常会話音声」は非言語情報をモニタリングするだけで無視でき，音声情報だけで計算機に対する「処理対象音声」を検出（スポッティング）できる．たとえば，「え一，今日は何日☆」のように，言い淀んだ後に入力したい内容を故意に高く発声すれば，計算機が「今日は何日」の部分を認識し，その答えを教えてくれる．本機能は，人間同士の対面での会話をマイクでモニタリングして適用するだけでなく，電話での会話に対して適用しても効果的である．さらに本機能は問合せだけでなく，曲名を音声スポッタの形式で発話するとその曲が再生され，人間同士がそれに関して議論するような用途にも応用できる．

以下，2章において，提案する音声スポッタの利点を議論し，3章で具体的な実現手法を説明する．次に，4章で，音声スポッタの応用例として実現した「オンデマンド会話支援システム」と「BGM付き電話システム」の2つを紹介し，5章でその実装方法を述べる．そして，6章で評価実験とシステムの試用結果を述べ，音声スポッタの有効性を示す．最後に，7章でまとめと今後の課題を述べる．

☆ 本論文では，故意に高くした発声を文字の上の線で示す．

2. 音声スポッタ

音声スポッタとは，ユーザが通常の会話では口にしない不自然な発声をすることによって，自分の意思で処理対象音声を指定できる音声インタフェース機能である．本研究では，以下の2つの手順で発声した不自然な発話だけを処理対象音声と見なし，他の発話や各種雑音は処理対象外として無視する．

(1) 有声休止（「え一」や「あの一」等の任意の母音の引き延ばし）によって言い淀む．
(2) 言い淀んだ直後に，入力したい内容を故意に高い声で発声する．

人間同士の会話中で，「え一，今日は何日」と言い淀んだ後に通常の声の高さでいうことはあっても，「え一，今日は何日」のように高い声で言うことはほとんどない．そのため，つねに会話をモニタリングしていても，こうした有声休止と高い声の組合せを誤検出して会話を邪魔してしまう可能性が低い．

このように音声スポッタでは，会話中の各発話が音声認識されるかどうかを，ユーザが音声だけで意図的に制御できることが重要となる．従来の典型的な音声認識システムでは，処理対象音声と通常会話音声との識別ができなかったために，つねにマイクから音声が入力されている状況下では，ユーザの発声はすべて処理対象音声として扱われていた．そのため，音声対話システムや，音声認識に基づくCTI（Computer Telephony Integration）システムのように，ユーザとシステムが1対1で対話する形態が想定されており，人間同士の対面や電話での会話中には，音声認識を使い分けることはできなかった．そうしたことを可能にする他の方法として，従来のワードスポッティング技術を活用し，処理対象音声の直前につねに特定のキーワードを発声することをユーザに義務付けるアプローチも考えられる．たとえば，「コンピュータ」のような一般的な語や，「Casper」，「Maxwell」のようなシステム名をキーワードとしてスポッティングする方法である．しかし，人間同士の会話中にそうしたキーワードを口にするとユーザが意図しないときにシステムが誤作動してしまうため，口にしないように注意し続けなければならない．それに対して音声スポッタでは，有

話者 A「んー，今日って何日だったっけ？」
話者 B「あれ，何日だったっけ？じゃあ調べてみよう．えー，今日は何日（高い声で）．」

日付が画面表示や音声合成で提示される．

図 1 音声スポッタを用いた情報支援システム（オンデマンド会話支援システム）の使用例
Fig. 1 An example of using an information assistance system with the speech-spotter function.

話者 A「そろそろ BGM を変えようよ．」
話者 B「B'z はどうかな？」
話者 A「それだったら，この前ヒットした曲がいいな．」
話者 B「『今夜月の見える丘に』だね．」
話者 A「えー，今夜月の見える丘に（高い声で）．」

B'z というアーティストの『今夜月の見える丘に』の楽曲が再生される．

図 2 音声スポッタを用いたジュークボックスシステム（BGM 付き電話システム）の使用例
Fig. 2 An example of using a jukebox system with the speech-spotter function.

声休止と声の高さの 2 つの非言語情報の不自然な組合せ（人間同士の会話中に出現しない組合せ）をスポッティングすることで，そうした問題を解決できる．

音声スポッタでは，ユーザが音声認識システムにコマンド入力したいという意図を非言語情報で伝えられる点が新しいだけでなく，以下の 3 つの利点も得られる．

(1) 人間同士の会話中でも音声入力が可能
人間同士で会話を続けながら，ユーザの望むときだけ即座に音声認識システムを利用できる．たとえば，図 1 や 図 2 のように，人間同士が会話を行いながら，情報検索や BGM の変更等の機能を声だけで呼び出せる．

(2) マイク以外の入力デバイスが不要
ボタンやマウス，カメラといったマイク以外の入力デバイスを使わずに，音声だけで操作ができる．そのため，ユーザの手足の動作や，顔の向き，視線の動かし方等に制約がなく，電話での会話のように音声しか利用できない場面にも幅広く適用できる．ただし，出力デバイスに関しては音声以外も併用した方が効果的であり，たとえば 4 章で述べる応用例では，画面が利用できる際にはシステムの動作状況や選択肢の一覧を画面表示する．

(3) 不使用時の心理的負担がない
計算機の支援が不要な場合，ユーザはその存在を意識せずに振る舞うことができ，心理的な負担がない．上述したように特定のキーワードをスポッティングして支援する方式では，システムの誤作動を避けるため，そのキーワードを口にしないように注意する必要がある．音声スポッタでは，そうした「ある言葉を言ってはいけない」という禁句がなく，人間同士の会話中に任意の言葉を使うことができる．

3. 音声スポッタの実現方法

音声スポッタを実現するには，以下の 4 つの処理を逐次実行して非言語情報をモニタリングする必要がある．

(1) ユーザの音声から，言語情報に依存せずに有声休止を検出する．
(2) 検出した有声休止箇所を目印に，発話区間☆始端（音声認識開始点）を決定する．
(3) ユーザの発話内容から，自動的に発話区間終端（音声認識終了点）を決定する．
(4) 決定した発話区間中の音声が，高い声であるか通常の高さの声であるかを識別する．

3.1 有声休止区間の検出

有声休止区間の始端と終端の検出には，文献 9) のリアルタイム有声休止検出手法を採用する．この手法は，有声休止（母音の引き延ばし）が持つ 2 つの音響的特徴（基本周波数の変動が小さい，スペクトル包絡の変形が小さい）をボトムアップな信号処理によってリアルタイムに検出する．任意の母音の引き延ばしを言語非依存に検出できるという特長を持つ．

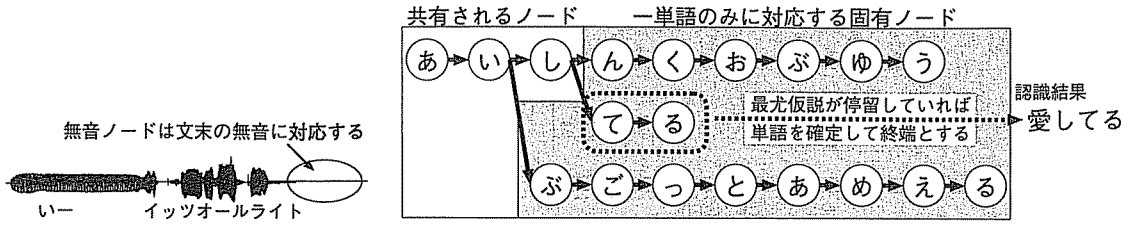
3.2 発話区間始端（音声認識開始点）の決定

文献 10) の発話区間の決定手法と同様に，検出した有声休止区間の終端に基づいて，発話区間始端を決定する（図 3）．単純には，検出した有声休止区間の終端を発話区間始端とすればよいが，有声休止で引き延ばした母音と同じ母音で処理対象音声が続く発話（たとえば，図 3 の「いー，イッツオールライト」等）での誤認識を避けるために，有声休止区間の終端から少し手前の時点を発話区間始端とする．具体的には，6 章の音声データとは異なる文献 10) の音声データに対する予備実験の結果から，130 ms 手前の時点とした．

3.3 発話区間終端（音声認識終了点）の決定

発話区間始端は有声休止によってユーザが明示的に指示できるとよいが，終端の方は自動的に決定された

☆ ここでの発話区間は，単に人が発話している区間を指すのではなく，音声認識システムが認識処理をする対象の区間を意味する．



(a) 一定時間無音ノードに停留した場合に音声認識を終了する

(b) 一定時間固有ノードに停留した場合に音声認識を終了する

図 4 発話区間終端（音声認識終了点）の決定
Fig. 4 Determining the end of an utterance.

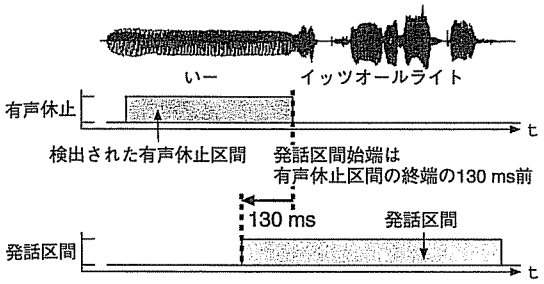


図 3 発話区間始端（音声認識開始点）の決定
Fig. 3 Determining the beginning of an utterance.

方がユーザの負担は少ない．そこで，内藤らの方法[11]と井ノ上らの方法[12]に基づいて，音声認識システムの各フレーム（10 ms）での最尤仮説が一定の条件を満たしたときに，発話区間終端であると決定する．音声認識システムは，単語認識のためのネットワーク文法（有声休止の母音から単語を経て無音へ遷移）を使用して，発話区間始端が決定された後に即座に認識処理を始める．ただし音声コマンドは，長いフレーズであっても 1 単語として木構造辞書に登録する．そして，毎フレームにおいて最尤仮説をチェックし，以下にあげる 2 つのノードのいずれかに一定時間以上（現在の実装では 200 ms 以上）停留していたら，そのフレームを発話区間終端とする．

(1) 無音ノードに停留している場合
最尤仮説が，文末の無音に対応した無音ノードに停留している場合，発話区間終端と見なす[11]．図 4 (a) のように，最尤仮説が連続して無音の状態であれば，入力中の音声が無音になった可能性が高く，発話が終わったと判断できる．

(2) 固有ノードに停留している場合
最尤仮説が，木構造辞書中で他の単語と共有されていない固有ノード（1 単語のみに対応し，他の単語の可能性がなくなったノード）に停留している場合，発話区間終端と見なす[12]．たとえば，図 4 (b) の 3 単語（アイシンクオブユウ，愛してる，アイブゴットアメエル）を持つ木構造辞書を探索中とする．ここでは，図中の網掛け部分が固有ノードに相当する．最尤仮説が破線で囲まれている「てる」に到達して停留し続けていれば，入力内容は「愛してる」である可能性が高く，発話が終わったと判断できる．

3.4 声の高さの識別

声の高低には個人差があるため，上記で決定した発話区間において意図的に声を高くしたかどうかは，絶対的な音の高さではなく，個人ごとに相対的に判断する必要がある．そこで，文献 13) の方法に基づいて，各発話区間の基本周波数（以下，F0）が，その個人の声の高さの基準となる基準基本周波数（以下，基準 F0）よりどれぐらい高いかで識別する．この基準 F0 は，有声休止区間中の F0 の平均として推定する[13]．有声休止中は，調音器官の変化が小さく[9]，F0 が安定する特徴を持つため，地声の F0（すなわち，基準 F0）に近くなる．最終的に，発話区間中の F0 の平均が，基準 F0 から相対的に見てある閾値以上高ければ，高く発声された「処理対象音声」であると識別する．

以上の処理の例を，図 5 に示す．図の左側は，「えー，今日の天気」「えー，何日.」を発話したときの F0 の変化の例である．図中，(A) の有声休止区間の平均 F0 が基準 F0 の推定に用いられる．(B) の発話区間の平均 F0 は閾値未満であるため「通常会話音声」と識別され，(C) の発話区間の平均 F0 は閾値以上であるため「処理対象音声」と識別される．この (C) の音声認識結果だけを音声コマンドであると解釈することで，音声スポッタの機能は実現される．

4. 音声スポッタの応用システム

音声認識技術をインタフェースとして活用する際の新たな応用例を提示するとともに，音声スポッタの有効性を確認するために，「オンデマンド会話支援システム」と「BGM 付き電話システム」の 2 つの応用シ

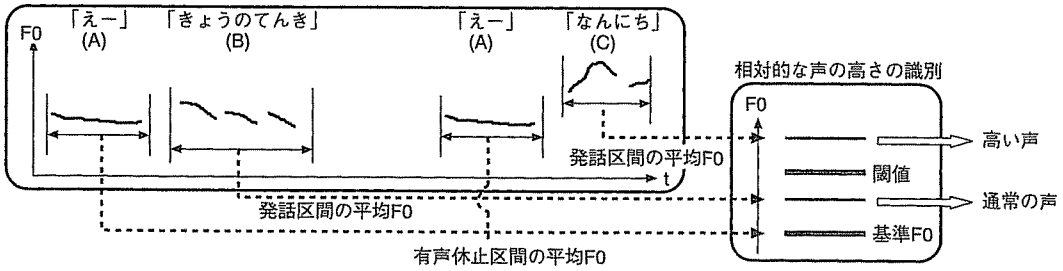


図 5 声の高さの識別処理の例：「えー，今日の天気」「えー，何日」（「何日」は故意に高く発声）を発話したときに，(A) の有声休止区間の平均 F0 から基準 F0 が推定される．(B) の平均 F0 は閾値未満であるため通常の声と識別され，(C) の平均 F0 は閾値以上であるため高い声と識別される

Fig. 5 An example of judging the voice pitch.

ステム☆を提案する．

4.1 オンデマンド会話支援システム

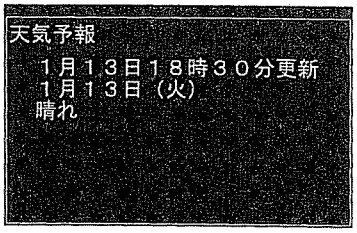
「オンデマンド会話支援システム」は，図 1 のように，人間（ユーザ）同士の会話中に，ユーザが望むときだけ，音声で必要な情報を検索できるシステムである．情報検索結果は画面表示または音声合成のどちらか一方，もしくは両方へ出力できる．

図 6 に，オンデマンド会話支援システムの実行画面の例を示す．システムが起動すると，情報表示用のウィンドウが開く．ユーザが情報支援を受けたいときには，まず，音声スポッタの発声方法に従い，故意に母音を延ばして言い淀む．すると，画面上部に有声休止らしさを表す緑色のバーが左端に表示され，母音を延ばしている間，左から右に向かって延びていく．そのまま言い淀み続けて右端に到達すると，システムが有声休止を検出してバーが赤色になる（図 6 (a)）．このように，ユーザがどれぐらい母音を延ばせば有声休止として検出されるのかが，視覚的に分かるようになっている．次にユーザは，事前に決められた様々な音声コマンドの中の 1 つを高い声で発声する．すると，「処理対象音声」であると識別したことを示す効果音が鳴り，ユーザの求める情報が画面に表示される（図 6 (b)）．

4.2 BGM 付き電話システム

「BGM 付き電話システム」は，図 2 のように，人間（ユーザ）同士が電話で通話をしながら，ユーザの望むときに，楽曲の選曲・再生が可能なシステムである．電話の受話器は音声入出力装置（マイクとスピーカ）を標準搭載しており，音声認識技術の応用先として魅力的であるが，従来の電話を活用した CTI システムやボイスポータルシステムでは，人間同士の会話

(a) 母音の引き延ばしを検出し始めると，画面上部に表示される緑色のバーが左端から徐々に長くなっていき，さらに言い淀み続けてバーが右端に達して赤くなると，「処理対象音声」の入力が可能となる



(b) 言い淀んだ直後に高い声で「今日の天気」と発声すると，システムからその回答が得られる

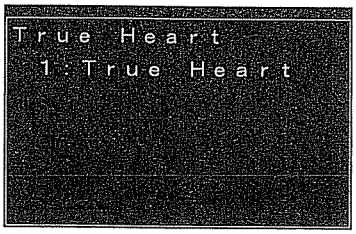
図 6 オンデマンド会話支援システムの画面表示例

Fig. 6 Screenshots of the on-demand human-human conversation support system.

を支援することは考えられていなかった．ここではその支援の一例として，ユーザが遠隔地間で電話による通話をしている最中に，あたかも 2 人が一緒の部屋にいて BGM を流しながら会話を楽しんでいるかのように，同じ BGM を聞きながら会話ができるシステムを実現した．

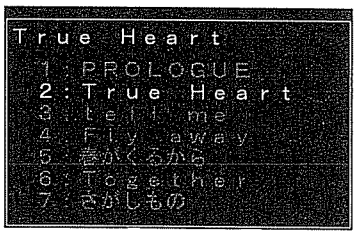
図 7 に，BGM 付き電話システムの実行画面の例を示す．図 6 (a) のオンデマンド会話支援システムと同様に，ユーザが母音を引き延ばして有声休止を発声すると，画面にフィードバックが表示される．選曲の候補や結果は，画面出力だけでなく，音声合成出力も可能となっている．本システムでは，以下の 2 つの方法で音楽を聞くことができる．

☆ 2 つの応用システムの動作の様子のビデオは，WWW サーチエンジン等で，“Speech Spotter”，“on-demand information system”，“music-playback system” の 3 つを含むページを検索するとアクセスできる．



(a) 曲名から音楽再生する例：「えー，TrueHeart」のように，有声休止後に高い声で曲名を発話すると，その曲が再生される

(b-1) アーティスト名から選曲して音楽再生する例：「んー，緒方智美」のように，有声休止後に高い声でアーティスト名を発話すると，緒方智美の曲目一覧が表示される



(b-2) アーティスト名から選曲して音楽再生する例（つづき）：曲目一覧を見た後に，「えー，2番」のように再び有声休止後に高い声で選択したい曲の番号を発話すると，その曲が再生される

図7 BGM付き電話システムの画面表示例
Fig. 7 Screenshots of the telephone system with BGM-playback function.

(1) 曲名を言って楽曲を再生する方法
音声スポッタの発声方法に従い，「えー，TrueHeart」のように有声休止の後に曲名を高い声で発声すると，高い声と識別したことを示す効果音が鳴り，その曲名の楽曲が流れ始める．その際，画面出力には図7(a)のように曲名が表示され，音声合成出力では曲名を読み上げる．

(2) アーティスト名を言って楽曲一覧から選曲して再生する方法
「んー，緒方智美」のように有声休止の後にアーティスト名を高い声で発声すると，効果音が鳴り，図7(b-1)のようにアーティスト名とそのアーティストの曲目の一覧が表示される．音声合成出力では，「緒方智美の曲は，1番プロローグ，2番トゥルーハート，...」のように，そのアーティスト名と曲目を順に読み上げていく．そして，ユーザがそれらの中から希望楽曲の番号か曲名を「えー，2番」のように発話すると，その楽曲が流れ始める．ユーザは，システムの音声出力中に割り込んで音声入力して選択できる．選択すると，画面上では図7(b-2)のようにその曲名の色がハイライトされ，合成音声では曲名を読み上げる．

本システムの特長として，音楽再生中や音声合成中でも，音声コマンド入力や人間同士の自由な会話が可能な点があげられる，たとえば，楽曲再生中に，「んー，ストップ」等を入力して再生を停止したり，別の曲名を言って楽曲を変更したりすることができる．また，会話のBGMとして楽曲を再生するだけでなく，実際に楽曲を聞きながらその曲について議論したり，感想を述べ合うような使い方ができる．このようなBGMの共有や再生楽曲についての意見交換は，従来遠隔地間では非常に困難だったことであり，特に音楽を日常的に自らの意志で聞く人々に対して有効である．さらに，このシステムを4.1節のオンデマンド会話支援システムと組み合わせ，人間同士が実際に一緒にいる部屋で，BGMを変える目的で使用してもよい．

5. 音声スポッタを組み込んだ音声インタフェースの実装

「オンデマンド会話支援システム」と「BGM付き電話システム」は，基本となる音声スポッタ機能を利用し，共通のシステム構成要素で実現できる．以下では，音声スポッタ機能の共通部分の実装について述べた後に，各応用システムの実装の違いを述べる．

5.1 音声スポッタの実装

図8に，音声スポッタ機能を構成する各システム構成要素（プロセス）と，全体の処理の流れを示す．プロセスは図中の囲み字で示されており，ネットワーク（LAN）上の複数の計算機で分散して実行することができる．プロセス間の通信には，音声言語情報をネットワーク上で効率良く共有することを可能にするネットワークプロトコルRVCP（Remote Voice Control Protocol）[14]を用いた．

処理の流れについて説明する．まず，マイク等から音声入力部（Audio signal input）に入力された音響信号は，ネットワーク上にパケットとして送信される．特徴量抽出部（Feature extractor），有声休止検出部（Filled-pause detector），F0推定部（F0 estimator）がそのパケットを同時に受信し，並行して処理する．次に，発話区間検出部（Endpoint detector）は，有声休止区間の終端情報を有声休止検出部から受け取って

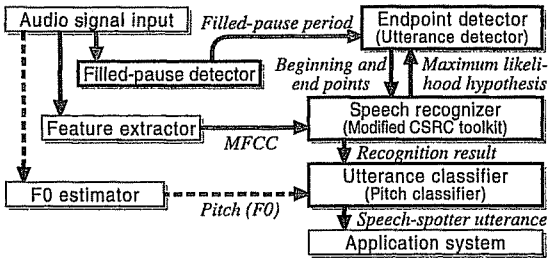


図 8 全体の処理の流れ
Fig. 8 System architecture.

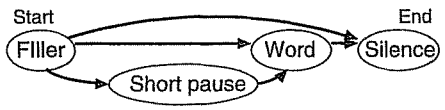


図 9 音声スポッタを組み込んだシステムで用いる文法
Fig. 9 Grammar for speech-spotter systems.

発話始端を決定し，音声認識部（Speech recognizer）にその情報を送信する．音声認識部は特徴量抽出部から MFCC（mel-frequency cepstral coefficients）パラメータを受け取り，検出された発話始端から認識処理を開始する．そして，最尤仮説を毎フレームごとに発話区間検出部に送る．発話区間検出部は受け取った最尤仮説をもとに発話終端を決定し，その結果を音声認識部に送信する．音声認識部が発話終端時点の認識結果を声の高さの識別部（Utterance classifier）へ送ると，声の高さの識別部は，並行して受信した F0 推定部の結果に基づいて，発話が通常の声か高い声かを識別する．最後に，高い声の発話と識別された認識結果だけが，応用システム（Application system）に通知され，認識結果中の音声コマンドが実行される．

音声認識部は，発話終端を決定するために，毎フレームの最尤仮説を発話区間検出部に送信できる必要がある．そこで，CSRC の日本語ディクテーション基本ソフトウェア（julian 3.3beta）[15),16)] を RVCP に対応させ，最尤仮説の送信が可能なように拡張して用いた．音声特徴量は，MFCC 12 次元 + ΔMFCC 12 次元 + Δpower 1 次元の計 25 次元とした．音響モデルは，ASJ-JNAS，ASJ-PB の男性話者 133 人分（計 20,414 文）[17)] から学習し，混合数 16，状態数 2,000 のトライフォンモデルとした．今回は単語認識を対象とし，図 9 に示すネットワーク文法を使用する．この文法では，有声休止を含むつなぎ語（Filler）から認識対象となる単語（Word）に直接遷移するか，短い無音（Short pause）を経て遷移し，最終的に無音（Silence）に到達する．ただし，つなぎ語としては有声休止以降を登録する．たとえば，「あのー」は末尾の母音の「おー」の部分だけを登録する．

5.2 オンデマンド会話支援システムの実装

図 8 の応用システム（Application system）の部分を変更することで，以下の機能を持つオンデマンド会話支援システムを実装した．

- 天気情報の検索機能
  「あー，明日の天気」のような天気に関する問合せに対して，天気情報（今日，明日，明後日等）を提示する．天気情報を記載している WWW ページから，http 経由でテキストデータとして取得して実装した．
- 現在の日時の照会機能
  「えー，今日は何日」のような日時に関する問合せに対して，現在の日時を提示する．システムが動作する OS（Linux）上の時刻を取得して実装した．時刻は，Network Time Protocol（NTP）で同期しておく．
- ニュースの見出しの表示機能
  「んー，経済ニュース」のようなニュースに関する問合せに対して，最新の各種ニュース（経済，国際，政治，社会，スポーツ等）の見出し部分を表示する．ニュースを記載している WWW ページから天気情報と同様に取得して実装した．

上記以外の情報支援にも，音声認識結果を解釈する新たなプロセスを追加するだけで，容易に対応できる．

5.3 BGM 付き電話システムの実装

図 8 の応用システムの部分を変更し，楽曲再生機能を組み込むことで，BGM 付き電話システムを実装した．楽曲を管理するためのデータベースサーバとしては SQL サーバを使用し，アーティスト名とその曲名，それらの曲に対応するサウンドファイル名を事前に登録しておく．さらに，アーティスト名と曲名は，それぞれ，音声認識システムの単語辞書上（言語モデル上）の 1 単語として登録する．楽曲データベースとしては，「RWC 研究用音楽データベース：ポピュラー音楽」[18)] の全 100 曲（RWC-MDB-P-2001 No.1〜100）を使用し，楽曲名 100 語，アーティスト名 34 語を登録した．

なお現在の実装では，電話の受話器のマイク経由の音声ではなく，別のヘッドセットマイクからの音声を認識する形態とした．

6. 評価実験

音声スポッタの基本性能として，会話相手に対する「通常会話音声」を計算機に対する「処理対象音声」と誤検出しないことと，音声スポッタによる発声を適切に「処理対象音声」として検出できることを確認する．

表 1 実験結果：通常会話音声を処理対象音声と誤って識別した回数
Table 1 Experimental results: the number of mistakes where conversational utterances are detected as command utterances.

	(a) 声の高さだけ	(b) 音声スポッタ
誤識別回数	1,338 回	60 回

また，音声スポッタを組み込んだ 2 つの応用システムの試用結果を述べる．

6.1 音声スポッタの基本性能の評価

音声スポッタでは，有声休止と声の高さの 2 つによって，通常会話音声と処理対象音声（音声スポッタの発声）を識別するが，声の高さのみを用いて識別する簡易な手法も考えられる．そこで，

(a) 声の高さだけで識別

(b) 有声休止と声の高さで識別（音声スポッタ）

の 2 つの手法で識別した性能を比較評価し，音声スポッタの有効性を確認する．本評価では，(a) と (b) を的確に比較するために，声の高さの識別に用いる閾値は，双方共通の値とする．その閾値は，文献 13) の研究の経験から 300 cent（平均律の半音 3 個分の音程）とした．

6.1.1 通常会話音声の棄却性能評価実験

通常会話音声を処理対象音声と誤って識別した回数が少ないほど，人間同士の会話に不必要に干渉しない優れた手法といえる．そこで，PASD 音声コーパス（重点領域模擬対話音声コーパス）[19] を対象に，手法 (a) と手法 (b) のそれぞれで処理対象音声を検出し，これを検証する．ただし，手法 (b) では声の高さの判断基準（基準 F0）は検出した有声休止から決まるが，手法 (a) での基準 F0 は事前に与える必要がある．そのために，各データから事前に有声休止区間だけを自動検出し，基準 F0 を求めて手法 (a) で用いることとした．このように有声休止区間は実験に不可欠なので，PASD 音声コーパスの中で，有声休止区間の存在する約 6 時間分のデータ（81 データ）を対象に実験した．これらは人間同士の通常の対話であるため，処理対象音声は含まれておらず，すべて通常会話音声に相当する．

実験結果を表 **1** に示す．この比較結果から，手法 (a) の声の高さだけで識別しようとする方が誤りが多いことが分かる．音声スポッタによって，有声休止が先行することを条件に加えることで，誤りを大幅に (95.5%) 削減できた．この結果から，音声スポッタで 2 つの非言語情報を組み合わせることの有効性が確認された．なお，2 章で議論したようなワードスポッティング技術でキーワード検出して支援する方法の場合には，上記のような誤識別回数は，「通常会話音声中にキーワードがシステムへの入力を意図せずに何回発声されたか」に依存する．つまり，適切に動作するかどうかは，通常会話中のユーザの意識的な振舞いに委ねられており，そうしたキーワードが人間同士の会話での禁句となって，ユーザに心理的負担を与えてしまう．2 章の末尾でも述べたように，音声スポッタの非言語情報の組合せは，そうした禁句を作らない点でも有効である．

表 2 実験結果：処理対象音声の検出結果
Table 2 Experimental results: detection performance for command utterances.

	(a) 声の高さだけ	(b) 音声スポッタ
正答	170 個	170 個
置換誤り	42 個	25 個
脱落誤り	6 個	23 個
挿入誤り	265 個	3 個
再現率	0.78	0.78
適合率	0.36	0.86
F 値	0.49	0.82

6.1.2 音声スポッタの発声の検出性能評価実験

音声スポッタの発声を適切に検出できることを確認するために，音声スポッタの発声を含む音声データを対象に，手法 (a) と手法 (b) のそれぞれで処理対象音声の検出実験を行う．そのために，音声認識を専門とする大学研究室に所属する 20 代男子学生 12 名から，音声スポッタの発声をした単語 218 語を含む約 40 分の音声データを収録した．収録では，事前に用意した文書を自発的に話しているように読み上げることとしたが，その文章中で，218 カ所の単語の部分は音声スポッタの特殊な発声をするように印を付けた．音声スポッタの発声をする場所以外では，咳払いや言い淀み等を許可し，できるだけ自然に読み上げた音声を収録することを目指した．

実験結果を表 **2** に示す．この結果から，手法 (a) と手法 (b) の両者の正答数（処理対象音声を処理対象音声であると正しく識別した個数）は同じであるが，手法 (b) の方が，手法 (a) に比べて挿入誤り（通常発声音声を処理対象音声であると誤識別した個数）が大幅に減少しており，優れていることが分かる．つまり，音声スポッタでは再現率を低下させることなく適合率を上げることができ，誤りを大きく削減できたことが確認された．

6.2 音声スポッタの試用結果

音声スポッタを組み込んだ 2 つの応用システムを，音声認識を専門とする大学研究室に所属する 20 代男子学生 6 名が，主に居室での雑談中に数日間試用した．いずれも研究室内での発表会において，音声スポッタの機能と動作原理はすでに把握していた．試用の結果，

音声スポッタを用いることで，特別な訓練をしなくても，数回試した後には人間同士の自由な会話中に音声認識システムを利用することができた．また，システムの支援が不要なときに，わざわざマイクのスイッチを切らなくてもよく，必要なときだけ音声スポッタの不自然な発声をすれば支援が得られるため，手間がかからず有用であった．実際に使ってみると，音声スポッタの発声の検出性能は，6.1 節で示した性能よりも体感上は高く感じることが多かった．これは，有声休止で母音を引き延ばす際に，どれぐらい延ばせば検出されるのかが画面表示による視覚的なフィードバックによって分かるためである．

6.2.1 オンデマンド会話支援システムの試用結果

2 人の人間がマイクの近くで会話中に，音声で，天気や日時，ニュースを調べることができ，本システムが的確に動作することが確認された．従来は会話中にそれらの情報を知りたくなると，WWW ブラウザ等を操作して調べるわずらわしさがあったが，本システムでは音声入力だけで手軽に確認ができ，便利であった．通常と異なる発声方法のときだけシステムが反応すること自体を面白いと感じるユーザもおり，普通の発声と音声スポッタの発声とを，会話中に遊び感覚で使い分ける場面も見られた．

6.2.2 BGM 付き電話システムの試用結果

音声スポッタを使うことで，市販の携帯電話や固定電話で通話をしながら，ユーザが好きなときに BGM を鳴らすことができた．携帯電話自体から音楽を再生する着信メロディーや待ち受けメロディーはすでに広く使われているが，通話中に BGM を聞くという感覚は従来経験したことのないものであり，試用したユーザに好評であった．なお，音楽は携帯電話の音声と同じ伝送路を通るため，実験に用いた様々な携帯電話のキャリアや機種によって，音質が大きく異なることが分かった．

7. おわりに

本論文では，故意に言い淀んでから故意に高い声で発声するという不自然な発声をあえて採用することで，ユーザが処理対象音声と，そうでない通常会話音声とを使い分けることを可能にする「音声スポッタ」という音声インタフェース機能を提案した．さらに，音声スポッタを有効に活用する 2 つの応用システムを提案した．1 つ目の「オンデマンド会話支援システム」では，人間同士で会話中に，計算機による情報支援を音声だけで受けることを可能にした．2 つ目の「BGM 付き電話システム」では，自宅で BGM を流しながら友人と会話する行為を，携帯電話を用いて遠隔地間で行うことを可能にした．これは，我々の調査した限り，人間同士の電話による会話中に，話者が音声だけで計算機の支援を意図的に得ることを初めて可能にした事例である．これらのシステムを試用した結果，日常会話をしながら実際に音声認識システムを利用できることを確認した．

本研究は，音声インタフェースでの非言語情報の新しい活用法を切り開くことで，音声の持つ潜在能力を引き出すことを目指した「音声補完」[14),20)]，「音声シフト」[13)]，「音声スタータ」[10)] の一連の「音声補完シリーズ」研究の第 4 弾に位置付けられる．従来の音声インタフェース研究の多くが自然な発声を主眼に置いていたのに対し，音声スポッタでは，不自然さを意図的に利用することで便利な機能が実現できることを示し，音声インタフェースの新たな可能性を広げることができた．今後も非言語情報を積極的に活用することで，音声の潜在能力を引き出す多様な音声インタフェース機能を実現していく予定である．

参考文献

1) 河原達也，石塚健太郎，堂下修司：発話検証に基づく音声操作プロジェクタとそれによる講演の自動ハイパーテキスト化，情報処理学会論文誌，Vol.40, No.4, pp.1491–1498 (1999).
2) Rohlicek, J.R., Russell, W., Roukos, S. and Gish, H.: Continuous hidden Markov modeling for speaker-independent word spotting, *Proc. ICASSP 89*, pp.627–630 (1989).
3) Kawahara, T., Ishizuka, K., Doshita, S. and Lee, C.-H.: Speaking-style dependent lexicalized filler model for key-phrase detection and verification, *Proc. ICSLP 98*, pp.3253–3256 (1998).
4) Méliani, R.E. and O'Shaughnessy, D.: Powerful syllabic fillers for general-task keyword-spotting and unlimited-vocabulary continuous-speech recognition, *Proc.ICSLP 98*, pp.811–814 (1998).
5) Lyons, K., Skeels, C., Starner, T., Snoeck, C.M., Wong, B.A. and Ashbrook, D.: Augmenting Conversations Using Dual-Purpose Speech, *Proc. UIST 2004*, pp.237–246 (2004).
6) Murai, K., Kumatani, K. and Nakamura, S.: A robust end point detection by speaker's facial motion, *International Workshop on Hands-Free Speech Communication (HSC 2001)*, pp.199–202 (2001).
7) 松坂要佐，東條剛史，小林哲則：グループ会話に参与する対話ロボットの構築，信学論 (D-II)，

Vol.J84-D-II, No.6, pp.898–908 (2001).
8) Nagao, K. and Takeuchi, A.: Social Interaction: Multimodal Conversation with Social Agents, *Proc. AAAI-94*, pp.22–28 (1994).
9) 後藤真孝，伊藤克亘，速水 悟：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，信学論（D-II），Vol.J83-D-II, No.11, pp.2330–2340 (2000).
10) 北山広治，後藤真孝，伊藤克亘，小林哲則：音声スタータ："SWITCH" on Speech，情報処理学会研究報告音声言語情報処理 2003-SLP-46-12, pp.67–72 (2003).
11) 内藤正樹，黒岩眞吾，山本誠一，武田一哉：部分文仮説のゆう度を用いた連続音声認識のための音声区間検出法，信学論（D-II），Vol.J80-D-II, No.11, pp.2895–2903 (1997).
12) 井ノ上直己，中村 誠，酒寄信一，山本誠一，谷戸文廣：単語固有セルでのゆう度判定を用いた音声認識処理の高速化手法，信学論（D-II），Vol.J79-D-II, No.12, pp.2110–2116 (1996).
13) 尾本幸宏，後藤真孝，伊藤克亘，小林哲則：音声シフト：音高の意図的な変化を利用した音声入力インタフェース，信学論（D-II），Vol.J88-D-II, No.3, pp.469–479 (2005).
14) 後藤真孝，伊藤克亘，秋葉友良，速水 悟：音声補完：音声入力インタフェースへの新しいモダリティの導入，コンピュータソフトウェア（日本ソフトウェア科学会論文誌），Vol.19, No.4, pp.10–21 (2002).
15) 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄：IT Text 音声認識システム，オーム社 (2001).
16) Lee, A., Kawahara, T. and Shikano, K.: Julius — an open source real-time large vocabulary recognition engine, *Proc. Eurospeech 2001*, pp.1691–1694 (2001).
17) Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus, *Proc. ICSLP 98*, pp.3261–3264 (1998).
18) 後藤真孝，橋口博樹，西村拓一，岡 隆一：RWC 研究用音楽データベース：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌，Vol.45, No.3, pp.728–738 (2004).
19) 板橋秀一：PASD コーパス — 重点領域模擬対話音声コーパス，音声による人間と機械の対話，オーム社，pp.361–375 (1998).
20) 後藤真孝：解説 "音声補完：言い淀むと助けてくれる音声インタフェース"，情報処理（情報処理学会誌），Vol.43, No.11, pp.1210–1216 (2002).

(平成 18 年 6 月 21 日受付)
(平成 18 年 12 月 7 日採録)

後藤 真孝（正会員）

1993 年早稲田大学理工学部電子通信学科卒業．1998 年同大学大学院博士後期課程修了．同年電子技術総合研究所（2001 年に産業技術総合研究所に改組）に入所し，現在に至る．2000 年から 2003 年まで科学技術振興事業団さきがけ研究 21「情報と知」領域研究員，2005 年から筑波大学大学院助教授（連携大学院）を兼任．博士（工学）．音楽情報処理，音声言語情報処理等に興味を持つ．2000 年 WISS2000 論文賞・発表賞，2001 年日本音響学会粟屋潔学術奨励賞・ポスター賞，2003 年インタラクション 2003 ベストペーパー賞，2005 年情報処理学会論文賞等 19 件受賞．電子情報通信学会，日本音響学会，日本音楽知覚認知学会各会員．

北山 広治

2002 年早稲田大学理工学部電気電子情報工学科卒業．2004 年同大学大学院修士課程修了．同年（株）東芝入社．大学では音声インタフェースの研究に従事．東芝では動画コーデック LSI の研究開発を担当．

伊藤 克亘（正会員）

博士（工学）．1993 年電子技術総合研究所入所．2003 年名古屋大学大学院情報科学研究科助教授．2006 年法政大学情報科学部教授．現在に至る．音声を主とした自然言語全般に興味を持つ．

小林 哲則（正会員）

1985 年早稲田大学大学院博士課程修了．工学博士．同年法政大学工学部電気工学科講師．同助教授を経て，1991 年早稲田大学理工学部電気工学科助教授．1997 年電気電子情報工学科教授．現在，コンピュータ・ネットワーク工学科教授．MIT，ATR，NHK 技研等の客員研究員を歴任．音声情報処理，動画像処理等知覚情報システムの基礎研究およびその応用としての会話ロボットの研究に従事．2001 年度電子情報通信学会論文賞受賞．電子情報通信学会，日本音響学会，言語処理学会等の会員．