# 息継ぎ音を利用したコールセンター会話音声の発話分割

福田 隆　西村 雅史

# Speech Phrasing Based on Breath Detection for Telephone Conversations in Call Center

Takashi Fukuda and Masafumi Nishimura

近年，電話対応におけるコンプライアンス違反を早期発見・是正するマネージメント体制が重要視されている．その一環として，音声認識を利用した通話監視技術に注目が集まっている．コールセンターを対象とした音声認識では，会話音声をあらかじめ発話単位に分割し，認識処理の不要な無音部分を取り除いた後，検出された発話の単位で認識処理を行う．そのため，各発話は文脈上意味のある単位で区切られていることが望ましい．しかし，従来の発話検出技術は，雑音の混入した入力信号から人間の発声部分を正確に抽出することにのみ焦点が当てられており，発話の検出単位については検討されてこなかった．本論文では，人間の息継ぎ音（吸気音）に注目し，入力信号から吸気音を高精度に検出することによって，入力音声を文脈上意味のある単位に，自動に分割する方法を提案する．提案法では，呼吸音に特化した音響特徴量を利用し，識別器を段階的に構成することによって，吸気音を高精度に抽出する．提案法は 97.4% の吸気音検出精度を達成した．

This article discusses our research into the creation of a management system for call centers that, based on telephone conversations, detects and resolves compliance problems at an early stage. Automatic speech recognition (ASR) has been the subject of much attention over many years, with a view to employing it in call center frameworks. In the proposed ASR system which processes call center conversations, the system first divides the input signals into separate utterances and eliminates periods of silences, then feeds the utterances detected into the ASR engine. Doing so requires the splitting of the input signal into semantically grouped utterances. Conventional voice activity detection techniques have focused only on accurately extracting the speech segments from noisy signals. In our research, we focused on the breathing sounds during the telephone conversations and have attempted to split the input signals into grouped utterances of the proper lengths by extracting these sounds without the use of any language information. The proposed method leverages acoustic information specific to breathing sounds, leading to a two-step approach for the detection of speech phrases at an accuracy of 97.4%.

Key Words & Phrases：音声認識，コールセンター，モニタリング，発話分割，吸気音検出
Automatic speech recognition, Call center, Monitoring, Speech phrasing, Breath detection

## 1. はじめに

IBM が日本で初めて大語彙連続音声認識システム ViaVoice® を商品化して以来,10 数年が経過した［1］．これまでの音声認識は，キーボード入力の代替手段としての利用が中心であったが，耐雑音性の向上と共に，カーナビなどの車載機器や携帯端末（PDA: Personal Digital Assistant）で使われるようになった［2］．また近年では，認識処理が難しかった自由発話音声※1 の認識精度が向上してきたことから，コールセンターにおける通話記録の書き起こしや，会議議事録の作成支援など，音声認識利用の幅が拡大している［3，4，5］．

コールセンター業務における音声認識適用の背景の 1 つには，2007 年に完全施行となった日本版 SOX 法がある．これは相次ぐ会計不祥事やコンプライアンスの欠如を防ぐため金融庁が導入した金融商品取引法の一部であり，これに伴い，コンプライアンス強化を掲げる企業が増加した．その対策の一例が，電話営業活動の監視体制である．これまでは，監視役の人間が自身

提出日:2010年9月6日　再提出日:2010年11月26日

※1　人と人の会話のような自由発話音声では，言いよどみや，発音の変形（例えば「エヌ・エイチ・ケー」→「エネッチケー」），文法の逸脱などが発生するため，ニュースなどの原稿に沿って発話する読み上げ型の音声と比較して認識処理が難しい．

の経験に基づき，耳で聞いて内容を判断していたが，人手による通話の全数確認は事実上不可能であり，音声認識を利用した半自動監視技術に注目が集まっている．この技術は，通話内容を音声認識でテキスト化した後，禁止用語の検索やテキスト・マイニング技術と組み合わせて，不適切な電話対応を早期に発見しようという試みである．

大規模なコールセンターにおける全通話を音声認識の対象とする場合，1日当たり数千時間分もの音声データを処理しなければならないケースも存在する．これを実現するためには，分散処理技術の併用と共に，音声認識そのものの高速化が必要である．音声認識高速化の1つに，入力信号から人間が発話している部分をあらかじめ抽出しておき，その部分のみを音声認識の対象とする方法（すなわち，認識処理が不必要な無音区間をあらかじめ取り除いておく処理）がある［6］．入力信号から人間の発声部分だけを取り出す技術は発話区間検出（VAD：Voice Activity Detection）と呼ばれ，VADの適用により認識対象音声を大幅に削減することができる．

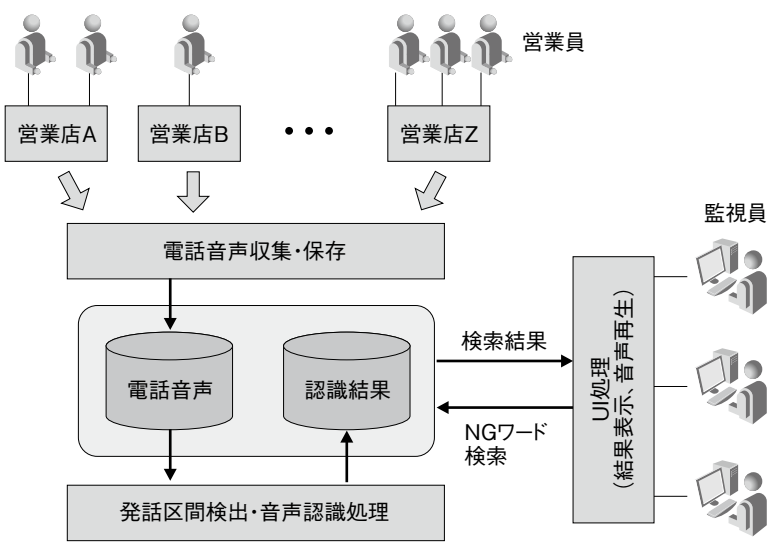


図 1. コール・モニタリング・システムの一例

通常，音声認識はVADシステムが検出した発話の単位で実行される．そのため，入力音声は人間にとって読みやすい単位（例えば，句読点に相当する位置）で区切られていることが望ましい.しかし,従来のVADは，雑音に埋もれた音声信号から，人間が発話している部分を漏れなく検出することに焦点が当てられており，また，その対象も音声コマンドなどの孤立単語（もしくは短文）発声であった［7］．そのため，標準的なVAD技術を雑音が少ない電話音声に適用すると，発話の検出漏れはほとんど発生しないが，複数の文が連結されて，過渡に長い発話として抽出されてしまうことがある．これは閾値（しきいち）の調整などで一部回避できるが，チューニングに頼った対処では，文脈上不適切なところで発話を分割してしまうという別の問題を引き起こす．適切な分割点を決定するタスクは，認識結果に対する可読性のみならず，音声認識性能にも影響を及ぼすため重要な問題である．

他方，人間は複数文を続けて発話する際，呼吸に伴う吸気行動を取るが，この吸気位置は文脈上の途切れ位置に相当するとの報告がある［8］．この事実はわれわれの感覚と合致する．本論文では，話者の息継ぎ音（吸気音）に着目し，吸気位置をVADにおける発話分割点として利用する方法を提案する※2．電話音声では，吸気音が明瞭なスペクトルとして観測できるため，このような利用が可能となる．観測信号から吸気音を抽出する研究は幾つか存在するが，自由発話音声を対象とした場合，抽出性能が不十分であり，雑音や子音などを吸気音と間違える湧き出し誤りが多発していた．本論文では，吸気音の音響的な特性と，吸気音の前後関係を考慮した高精度な吸気音検出法を提案する．

## 2. コール・モニタリング・システムの例と問題点

### 2.1 システムの概要

図1に音声認識を利用したコール・モニタリング・システムの一例を示す．各営業店の電話音声はネットワークを経由して音声データベースに保存され，蓄積された音声を順に音声認識処理によりテキスト化する．電話音声は発話区間検出技術により，人間の発声部分のみが切り出され，抽出された発話の単位で認識処理が行われる．認識結果はサーバーに保存され，時間情報と共に元の電話音声データと関係付けられる．監視員（モニタリング・チーム）は音声認識結果に対してキーワード検索を行い，検索結果の認識テキストと，それに対応する実際の音声を聞き比べながら不適切なコールを早期に発見する．

※2　直感的には，自分の知らない言語を音だけで判断して，適切な単位に分割するタスクと言える．

### 2.2 VADの観点から見た問題点

一般的なVADでは，まず，発話区間のスペクトル特性を表現する確率モデルと，非発話区間の特性を表現する確率モデルを学習データから事前に推定しておく．そして検出処理の段階では，入力音声がどちらのモデルに適合しているかを比較することによって発話区間を決定する．これを定式化すると次のようになる．

$$L(x_t) = \log p(x_t \mid M_1) - \log p(x_t \mid M_0) \quad (1)$$

ここで $\boldsymbol{x}_t$ は時刻 $t$ における音響特徴ベクトル※3である．また，$M_i$ は発話／非発話の特性を表現する確率モデル（$i$ = 0: 非発話，$i$ = 1: 発話）を表している．確率モデルとしては，ガウス混合モデル（GMM: Gaussian Mixture Model）が用いられることが多い．式（1）において，$L(\boldsymbol{x}_t)$ が閾値より大きい区間を発話，小さい区間を非発話区間と見なす．

VADにはさまざまな方式が提案されており，利用環境に対してチューニングを行うことで検出の単位（平均発話長）はある程度調整可能である．しかし，コールセンターのような自然発話音声を対象とする場合，以下に示す問題が発生する．

#### 1）分割に言語的誤りが生じるケース

過渡に長い発話区間が生成されるのを避ける最も単純な方法は，非発話区間に対する反応が鋭敏になるように，式（1）の閾値を調整することである．これにより，検出される発話長は平均して短くなるが，促音や自然発話特有の言いよどみ，フィラー※4などで発話が分割される別の問題を引き起こしてしまう．図2に具体例を示す．CASE 1は促音（「あさって」の「っ」の部分）で発話が分割されてしまう例である．促音は声道の閉鎖によって音が生成されるので，促音区間は無音に近い状態となる．そのため，声道の閉鎖時間が長い話者，すなわち無音状態が長くなるケースでは，発話継続中にもかかわらず，VADが非音声区間であると誤判断することが起こる．一方，CASE 2は音声の長音化に伴う分割の例である（「しーんよこはま（新横浜）」の「しーん」の部分）．これは発話者の場所（新横浜）に対する自信があいまいで，頭の中で確認しながら応答することによって，音素が長音化する例である．また，「しんよこはまーでしたっけ?」，「しんよこはまはー，最近きれいになりましたよねー」など，長音化の位置もまちまちであり，通常の長音（例えば「カレーライス」）などと比べて継続時間が長くなる傾向にある．長音化の部分は非音声のスペクトルと類似しており，促音のケースと同様，発話分割の対象位置になってしまう問題が生ずる．通常，これらの問題が顕在化することは少ないが，検出発話長が長くなるのを避けることと，言語的な分割誤りをなくすことはトレードオフの関係にある．

#### 2）検出発話長が長くなり過ぎるケース

VADが音声区間を重視，すなわち非音声区間に対しての反応が鈍い設計になっている場合，言語的に不適切な個所での発話分割はなくなるが，冒頭で取り上げた長い発話区間が抽出されてしまうという問題は依然として残る．図3に例を示す．熱心な営業努力の結果，一方的な発話となってしまったり，あるいは早口で発話した場合などには，文と文の間の無音区間が極端に短くなることがある．その結果，文の連結が発生し，図3のように営業員発話（Agent speech）が適切な個所で分割されず，一読しにくいテキストになってしまう．認識テキストを後から言語処理で分割するという方法も考えられるが，言語処理による分割ミスや処理量の増加は避けられない．また音声認識処理の性質上，音声認識への入力発話は長過ぎないことが好ましい※5．一方，コー

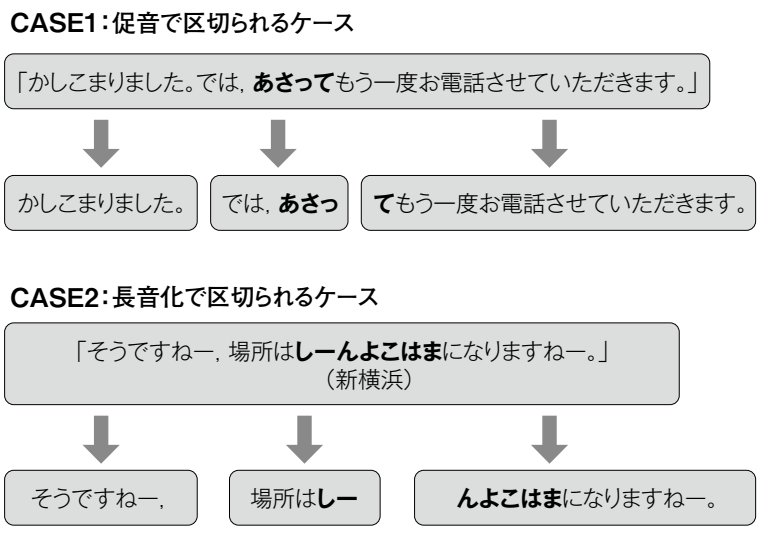


図2. 意味のない点で分割されるケース

※3　各時刻の信号成分をベクトルの形で表現したものである．通常，入力信号のスペクトル成分を特徴ベクトルとして利用する．
※4　「えーと」，「あのー」のような，発話と発話をつなぐために使われる単語．コール・モニタリングにおいては，フィラー自身も重要な情報になり得る．
※5　探索空間が広がるため，音声認識精度が低下する．

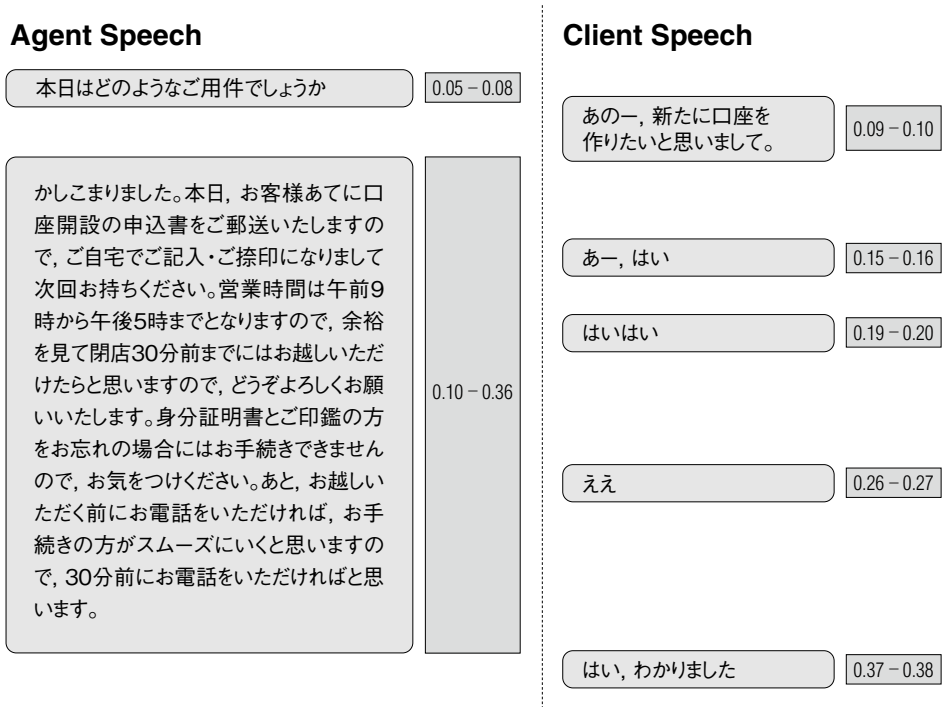


**図 3. 音声認識結果の表示例**

ル・マイニングにおいては，顧客（もしくは営業員）の相づち応答も重要な情報源となるが，検出発話長が長くなると，図 3 のように，どの発話に対する相づちであるかがあいまいなものとなってしまい，マイニング性能に影響を与えてしまう．

他方，検出発話長が長くなる別の原因として，電話音声特有の問題が存在する．受話器や携帯電話を耳に近づけて発話するタイプの電話音声（すなわち，ハンズフリー・フォンではない）では，人間の呼吸音がパワーの大きい信号として入力されるため，通常のマイクロホン環境では無音区間として見なされていた吸気音も，発話の一部としてとらえられてしまうことがあった．この場合も複数の文が連結され，1 つの長い発話区間として生成されてしまう．本論文では，この現象を踏まえて，明瞭なスペクトルとして観測される電話音声上の吸気音を，発話分割の情報として利用する．

## 3. 先行研究

### 3.1 発話分割

発話分割の先行技術として，特許公報［9］に入力信号を段階的に分割する方法が提案されている．この方法では，まず VAD によって入力信号を大ざっぱに発話区間に分割しておき，分割された発話について無音区間検出を実行した後，信頼性の高い無音区間での再分割を行う．分割の結果，所定時間以上の長さの発話については，さらに分割を繰り返す．最適な分割点が見つからない場合は，あらかじめ指定された時間もしくは文字数で強制的に分割する．一方，伊東らは少なくとも 2 チャネル以上で入力された対話音声について，相手チャネルで相づちが発生した時刻付近に，主チャネル側の発話分割に適した区間が含まれることを指摘している［10］．この結果を元に，相づち付近の無音区間を信頼して発話を分割する方法を提案している．しかしながら，上記 2 つの先行技術は，信頼性の高い無音区間や相づちなど，必ずしも所望の分割点に存在するわけではない情報を利用しているので，2 章で説明した問題点は依然として残る．

### 3.2 吸気音検出

息継ぎ音（吸気音）検出技術の先行研究は，音楽情報処理の分野で幾つか検討されている．吸気音の検出は，歌唱音声の収録において，吸気音を消したり強調したりする場面で有効であり，オーディオ編集ソフトに導入されている［11］．また，吸気位置は歌唱者のリズム感や間の取り方と関係があるという指摘があり［12］，さらに歌唱音声においては吸気音がフレーズ位置に相当するとの報告もある［13］．

一方，音声分析の分野では，Price らが GMM を用いた吸気音の識別方法を提案しており，93% の検出率

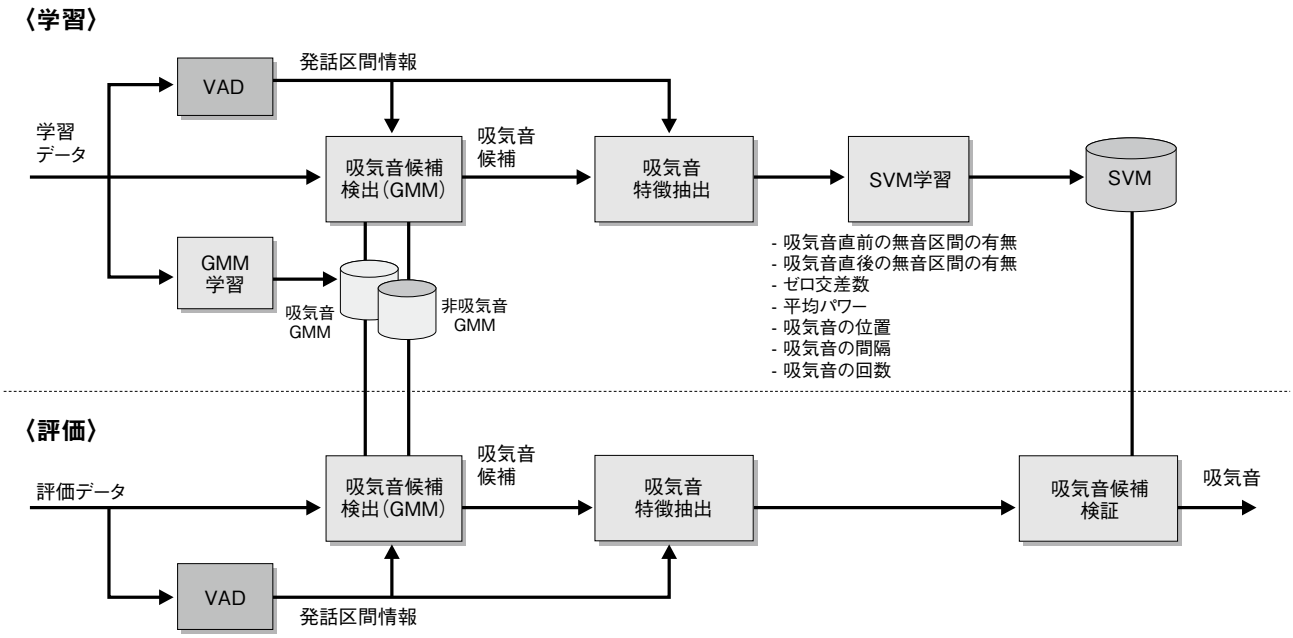


**図 4. 吸気音抽出処理の概要**

を得ている［14］．このほか，Wightman らは，スペクトル成分に対する離散フーリエ変換，もしくは離散コサイン変換から得られるケプストラム※6 を特徴量とし，ベイズ識別※7 を適用することによって最大で 91.3% の検出性能を得ている［15］．また同著者は，吸気音が文脈上の自然な区切りに位置するとも報告している［8］．吸気音はスペクトルの形状が雑音や子音の /s/ と類似していることから，雑音を吸気音と間違える湧き出し誤りが多発する．吸気音が検出できたかどうかのみに着目した検出率の観点で見れば，先行研究は高い性能を実現しているものの，湧き出し誤りも含めた指標（検出精度）でみると低い性能にとどまる※8．また，先行研究は歌唱音声や読み上げ音声を対象としており，一般的な会話などの自由発話音声で観測される多種多様な吸気音に対応できるわけではない．

## 4. 提案の吸気音検出法

### 4.1 処理の概要

本論文では，湧き出し誤りを低く抑えつつ，高い検出率を達成する吸気音検出法を提案する．図 4 にシステム構成を示す．提案システムでは，入力信号を標準的な VAD によってある程度の長さの発話区間に分割した後，個々の発話区間について吸気音検出を行うことによって，長い発話を細分化する．従って，発話区間に吸気音が存在しない場合は分割の対象とならないが，吸気音が存在しない文はもともとが短文であり，分割の必要がない．

提案手法では，最初に吸気音検出用の音響モデルを用いて吸気音の候補を検出し，その後，その候補が吸気音であるか否かを精査する 2 段階構成を採用する．この 2 段構成は，吸気音の候補を決定する段階において，あいまいなものも含めて可能性のあるものを漏れなく抽出し，それと同時に，雑音によって湧き出た候補を，後続の 2 クラス分類器で除外することによって，検出精度を高めることを狙っている．

音響モデルと 2 クラス分類器の学習段階では（図の上半分），まず，学習データを用いて，吸気音候補を抽出するための確率音響モデルを推定する．確率モデルとしては GMM を用いる．吸気音候補は入力音声に対して，吸気音 GMM と非吸気音 GMM との尤度（ゆうど）を比較することによって抽出する．定式化は式（1）と同様である．ここで抽出される情報は，何秒から何秒までが吸気音区間であるかというタイム・スタンプである．その後，得られた吸気音候補区間について，吸気音の特性を表す 7 つの要素（後述）を推定し，7 次元の特徴ベクトルを構成した後，2 クラス分類器を学習

※6　音声処理の分野で広く用いられている特徴量の 1 つ．スペクトル成分に対して，さらにスペクトル分析を行うことから，spectrum の最初の 4 文字を反転させて cepstrum（ケプストラム）と呼ばれている．
※7　音声処理を含むパターン認識の分野において代表的に用いられている方法．
※8　検出率と検出精度については 5 章で定義している．

する．2クラス分類器としては，サポート・ベクター・マシン（SVM: Support Vector Machine）を採用した※9．吸気音抽出段階（図の下半分）では，あらかじめ学習しておいたGMMを用いて吸気音候補を抽出した後，SVMによってその候補が吸気音であるかどうかを最終的に判断する．

### 4.2 吸気音特徴量

先行研究において，吸気音には「調波構造を持たない」，「継続時間が長い」，「吸気音の前後には極短時間の無音区間が存在する」，「子音 /s/ に比べてゼロ交差数が小さな値を取る」などの特性があることが示されている［16］．ここで調波構造とは，声帯振動に伴う基本周波数（声の高さに相当）と，その倍音成分から成るスペクトル平面上のしま模様を指す．また，ゼロ交差数とは，入力信号が音圧ゼロのレベルをまたぐ回数のことをいう．一方，吸気音の前後に一瞬の無音が存在するという事実は，実際に小説を朗読するなどして，自身の息継ぎのやり方を意識してみれば容易に確認されよう．これら先行研究の知見を踏まえて，本論文では「吸気音直前の無音の有無（無音が存在すれば1，存在しなければ0）」，「吸気音直後の無音の有無（存在すれば1，存在しなければ0）」，「ゼロ交差数」，「平均パワー」を表す素性を用いて4次元の特徴ベクトルを構成し，SVMで利用する．これら4つの音響的素性は個々の吸気音から独立に推定可能である．

本論文では，より一層の精度向上を目指して，吸気音のコンテキスト情報（時間的な位置情報）もSVMの入力素性として利用する．具体的には，a）発話区間と吸気音の位置関係，b）吸気音の間隔，c）発話区間内における吸気音の回数を用いる．ここで追加する3つの素性は，実際の電話音声から観測した統計情報に基づいている．a）について，発話区間の先頭，および終了付近に検出される吸気音は，雑音や子音 /s/ による湧き出し誤りである可能性が高い．そのため，発話区間を3つの区間に分け，吸気音候補がどの区間に属しているかを素性の値とする．具体的には，発話先頭付近= 0，発話終了付近= 1，それ以外= 2とした．b）の素性を導入する理由は，吸気音が短い時間内に連続して観測される場合，一方，もしくは両方が湧き出し誤りであることが多いためである．例えば，1秒以内に複数回の吸気音が観測されることは会話音声において不自然であるが，雑音などの影響によりこの現象が発生していた．最後にc）の素性について，吸気音の回数は発話長に依存するが，VADが出力する発話区間に対して最高でも5回程度であることを見いだした．しかし，従来手法ではこれを大きく超えて検出される例が散見されたので，この誤検出をc）の素性で吸収する．最終的に，これら3種類のコンテキスト情報に基づく素性と，先に示した4つの音響的素性を組み合わせて，7次元の特徴ベクトルを構成する．

## 5. 評価実験

コールセンターで収録された電話音声データを用いて，吸気音検出法の比較を行った．GMMとSVMの学習データには3.42時間分の電話音声を使用し，評価データには2.1時間分を用いた．VADには2.2節で説明したGMMに基づく方式を採用している．**表1**に実験結果を示す．表中の検出率，検出精度は次のように定義している．

$$\text{検出率}: \quad \frac{N_c}{N} \times 100 \tag{2}$$

$$\text{検出精度}: \quad \frac{N_c - N_f}{N} \times 100 \tag{3}$$

ここで，$N$はテスト・データ中に含まれる吸気音を人手でカウントした総数，$N_c$はシステムが検出した吸気音の内，人手で付与した吸気音位置と一致した数，$N_f$はシステムが吸気音として検出したものの，実際には吸気音ではなかった場合の数である．すなわち，$N_c + N_f$はシステムが（誤りを含めて）吸気音であると判定したものの総数であると見なすことができる．

表の従来方式はGMMに基づく方法を表しており，提案システムの前段（GMMを用いて吸気音候補を出力する部分）と同様である．ただし，提案システムでは

**表1. 実験結果**

| | 検出率（%） | 検出精度（%） |
|---|---|---|
| 従来方式（GMM） | 99.8 | 36.1 |
| 提案手法A（音響情報のみ） | 98.6 | 95.7 |
| 提案手法B（音響＋コンテキスト情報） | 98.9 | 97.4 |

※9 サポート・ベクター・マシンは，入力データがどちらのクラスに属するかを決定する問題において，優れた性能を示す機械学習の1つとして知られている．例えば，身長，体重，ウエスト・サイズの3要素を特徴ベクトルとして入力し，それが男性のものか，女性のものかを判断する問題などを扱うことができる．本研究においては，GMMによって抽出された吸気音候補が実際に吸気音であるか，それ以外であるかを決定するために利用する．


**図 5. 音声認識結果の表示例**

可能性のある候補をできるだけ多く抽出するように設計しているのに対し，**表1**の従来手法では，高い検出率を保ちつつ，検出精度が最も高くなるようにチューニングを行った．提案手法 A は SVM のための特徴ベクトルに音響情報のみ（4 次元）を用いた場合，そして提案手法 B は吸気音のコンテキスト情報も含めた 7 つの素性すべてを利用した場合の結果である．

まず，従来の吸気音検出方式（GMM）と提案手法 A，B を比較すると，検出率に関してはほとんど差がない．しかし，検出精度について見れば，従来手法はとても低い値を示している．これはシステムが吸気音であると誤認識した数 $N_f$ が極めて多いためである．一方，提案手法は顕著な改善を示しており，SVM による吸気音候補の絞り込み効果が大きい，すなわち湧き出し誤り $N_f$ を少なく抑えていることが分かる．さらに，2 段階構成の方式内（提案手法A，B）で比較すると，特徴ベクトルに吸気音のコンテキスト情報を加えることで，精度が 95.7% から 97.4% に向上しており，SVM の入力特徴ベクトルに音響情報のみを用いた提案手法 A と比較して，相対的に約 40% の誤り削減を達成している．発話分割に吸気音情報を利用するという目的においては，検出精度が高い（すなわち，湧き出し誤りがない）ことが重要である．その観点で見ると，従来手法は湧き出し誤りが極めて多いため，これを発話分割に利用した場合，単語の途中など，予期せぬ個所での分割が行われ，可読性が大幅に低下する．これは音声認識処理にも大きな影響を及ぼす．一方，提案手法 B を見ると，湧き出し誤りはほとんどない．わずかに存在する湧き出し誤りは，発話開始前や終了直後の微小区間における背景雑音に起因するものであり，発話分割に影響を与えるものではない．

**表 2. 発話区間長の統計量（秒）**

| | 平均 | 標準偏差 | 最小 | 最大 |
|---|---|---|---|---|
| 従来方式（GMM） | 3.8 | 5.3 | 0.15 | 154.8 |
| 提案手法 B | 2.8 | 2.6 | 0.15 | 24.5 |

**図 5** に吸気音で長区間発話を分割した例を示す．吸気音情報を発話分割に利用することで，営業員発話の可読性が増すと同時に，顧客側の相づちとの関係が明瞭になっていることが分かる．人間の短期記憶にとって適度な発話長は 2.5 秒程度であるとの考察がある［9］．**表 2** に示すように，発話分割への吸気音の導入前後で,平均発話長は 3.8 秒から 2.8 秒に改善しており，そのような観点においても望ましい値になっている．また，

従来手法では約 154 秒もの長い発話区間が生成されていたのに対し，提案手法では最大でも 25 秒程度に収まることを確認した．

## 6. おわりに

本論文ではコールセンター会話音声から高精度に吸気音を検出する方法を提案し，吸気音を基点とした発話分割を行うことで，認識結果の可読性を向上させる方法を説明した．吸気音の検出は，候補の抽出と，それらの検証の 2 段階に分けることによって高い性能を実現している．検証用分類器（SVM）の入力特徴ベクトルとして，吸気音の音響特徴とコンテキスト情報の両方を用いることで 97.4% の検出精度を達成できることを示した．

音声認識の分野では，自由発話音声に対する性能改善と共に，感情分析に関する研究が始まっている．吸気音は間の取り方との関係があるとの考察があり，感情分析において吸気音情報が役に立つと期待している．また一方で，吸気音位置と相づちとの関係が議論されていることから，吸気音は対話音声システムにおける自動応答のタイミングとして最適な位置である可能性もある．今後は，このような応用研究に着手したい．

謝辞

提案手法について熱心に議論をいただいた日本アイ・ビー・エム株式会社 知的財産部門の安福満里氏，および東京基礎研究所の市川治氏，立花隆輝氏に深謝したい．

## 参考文献

［1］ 日本アイ・ビー・エム監修，“IBM ViaVoice 98,” インフォクリエイツ出版，(1998)．
［2］ 大淵康成，畑岡信夫：“音声認識技術の実用化に向けた自動車内実環境での評価実験,” 情報処理学会研究報告,2006-SLP-61 (1)，pp.1-6, (2006)．
［3］ 西村雅史：“音声認識ビジネスの現状と将来の展望,” 情報処理学会研究報告，2005-SLP-55 (3)，pp.13-15, (2005)．
［4］ H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig: “The IBM 2004 conversational telephony system for rich transcription,” *Proc. ICASSP*, pp. 205-208, (2005)．
［5］ G. Zweig, O. Siohan, G. Saon, and B. Ramabhadran, D. Povey, L. Mangu, and B. Kingsbury: “Automated quality monitoring in the call center with ASR and maximum entropy,” *Proc. ICASSP*, pp. 589-592, (2006)．
［6］ T. Fukuda, O. Ichikawa, and M. Nishimura: “Long-term Spectro-temporal and Static Harmonic Features for Voice Activity Detection,” *IEEE Journal of Selected Topics in Signal Processing*, Vol. 4, No. 5, pp. 834-844, (2010)．
［7］ 藤本雅清：“音声区間検出の基礎と最近の研究動向,” 電子情報通信学会，SP2010-23，pp.7-12，(2010)．
［8］ C. W. Wightman and M. Ostendorf: “Automatic Labeling of Prosodic Patterns,” *IEEE Trans., on Speech and Audio Processing*, Vol. 2, No. 4, pp. 469-481, (1994)．
［9］ 特開 2004-212799 号公報
［10］ 特開 2008-164647 号公報
［11］ Waves, “Waves | プラグイン | DeBreath,” <http://www.waves.com/content.aspx?id=2173>
［12］ 中野倫靖：“楽譜情報を用いない歌唱力自動評価手法,” 情報処理学会論文誌，Vol. 48，No. 1，pp. 227-236, (2007)．
［13］ 中村敏江：“音楽における「間」と呼吸について,” 日本音響学会音楽音響研究会資料，MA94-16，pp. 19-26, (1994)．
［14］ P. J. Price, M. Ostendorf, and C. W. Wightman: “*Prosody and Parsing,” Proc. Workshop on Speech and Language Processing*, pp. 5-11, (1989)．
［15］ C. W. Wightman and M. Ostendorf: “Automatic Recognition of Prosodic Phrases,” *Proc. ICASSP*, pp. 321-324, (1991)．
［16］ 中野倫靖，緒方淳，後藤真孝，平賀譲：“無伴奏歌唱におけるブレスの音響特性と自動検出,” 日本音響学会春季研究発表会講演論文集，1-11-12，pp. 265-268, (2008)．

**日本アイ・ビー・エム株式会社**
**IBM Research 東京基礎研究所**
**主任研究員**

福田 隆 Takashi Fukuda

［プロフィール］
2005 年，IBM 東京基礎研究所に入所以来，音声認識の研究に従事．2008 年より情報処理学会音声言語情報処理研究会運営委員．2010 年，日本音響学会粟屋潔学術奨励賞受賞．IEEE，日本音響学会，電子情報通信学会各会員．博士（工学）
fukuda1@jp.ibm.com

**日本アイ・ビー・エム株式会社**
**IBM Research 東京基礎研究所**
**主席研究員**

西村 雅史 Masafumi Nishimura

［プロフィール］
1983 年，日本 IBM 入社．以来東京基礎研究所において音声認識・音声合成の研究に従事．1998 年情報処理学会山下記念研究賞，1999 年，日本音響学会技術開発賞各受賞．電子情報通信学会英文論文誌編集幹事，シニア会員．IEEE，日本音響学会，情報処理学会各会員．音声技術担当．博士（工学）
nisimura@jp.ibm.com