社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（３）
2007／12／20

音響モデルと言語モデルに基づく音声区間検出を用いたハンズフリー音声認識アルゴリズムの評価

酒井　啓行†　ツィンツァレクトビアス†　川波　弘道†　猿渡　洋†　鹿野　清宏†
李　晃伸††

† 奈良先端科学技術大学院大学 情報科学研究科　〒630–0192 奈良県生駒市高山町 8916–5
†† 名古屋工業大学　〒466–8555 愛知県名古屋市昭和区御器所町
E-mail: †{hiroyuki-s,cincar-t,kawanami,sawatari,shikano}@is.naist.jp, ††ri@nitech.ac.jp

あらまし　人と音声対話ロボットとの自然な対話を可能にするためにハンズフリーインターフェースの導入が求められている．ハンズフリー音声認識システムでは様々な背景雑音の混入や，ユーザの直接音のパワーが減衰するなど様々な理由で入力音声の Signal-to-Noise Ratio (SNR) が低下してしまう．そして SNR の低下に伴いユーザの発話区間を特定する音声区間検出が困難となる，また雑音環境における有効な音声区間検出手法は確立されていない．本稿では，雑音環境下においても頑健にユーザの発話区間を検出する音響モデルと言語モデルに基づく認識による音声区間検出を用いたハンズフリー音声認識アルゴリズムの評価を行う．従来の VAD 手法として振幅パワー，統計モデル，GMM などに基づく手法を挙げ，性能比較実験を行うことで提案手法の有効性を示す．
キーワード　音響モデルと言語モデルに基づく認識による音声区間検出，ハンズフリー音声認識，実環境対話ロボット

Evaluation of Hands-free Speech Recognition Algorithm using Decoding Voice Activity Detection based on Acoustic and Language Models

Hiroyuki SAKAI†, Tobias CINCAREK†, Hiromichi KAWANAMI†, Hiroshi SARUWATARI†,
Kiyohiro SHIKANO†, and Akinobu LEE††

† Graduate School of Information Science, Nara Institute of Science and Technology　Takayama-cho 8916–5, Ikoma-shi, Nara, 630–0192 Japan
†† Nagoya Institute of Technology　Gokiso-cho, Showa-ku, Nagoya, Aichi, 466–8555 Japan
E-mail: †{hiroyuki-s,cincar-t,kawanami,sawatari,shikano}@is.naist.jp, ††ri@nitech.ac.jp

Abstract　Introduction of hands-free interface into speech recognition (SR) systems is expected for natural iteraction between humans and spoken dialogue robots. In hands-free SR system, Signal-to-Noise Ratio (SNR) of input signal becomes worse because of background noise in real-environment and other reasons. This will cause degradation in recognition performance when using conventional Voice Activity Detection (VAD). In this paper, we evaluate hands-free SR algorithm using decoding VAD based on acoustic and language models for robust VAD in noisy environment. We performed experiment for comparing proposed and conventional VAD method, for example, based on amplitude power, statistical model and GMM. And, we evaluate effectiveness of the proposed method.
Key words　Voice Activity Detection (VAD) by decoding based on Acoustic Model and Language Model, Hands-Free speech recognition, Real-environment spoken dialogue robot.

1. はじめに

近年音声認識の発達により，カーナビゲーションや音声情報案内や音声対話ロボット，携帯型音声翻訳機など実環境で利用される音声認識システムが登場している．多くの音声認識システムではヘッドセットやハンドマイク，指向性マイクなどを用いることでユーザの発声が際立って収録される，つまり入力音声の Signal-to-Noise Ratio(SNR) が高くなるため音声認識の

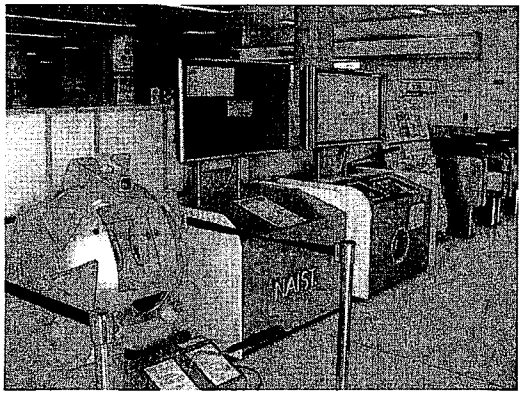

図 1 実環境音声路線案内システム "キタちゃん (奥), キタロボ (手前)": 2006 年 3 月より奈良県生駒市の学研北生駒駅に常設
Fig.1 Real-Environment Speech Train Information Guidance System "Kita-chan(back), Kitarobo(front)" : System installed at the local railway station since March 2006.

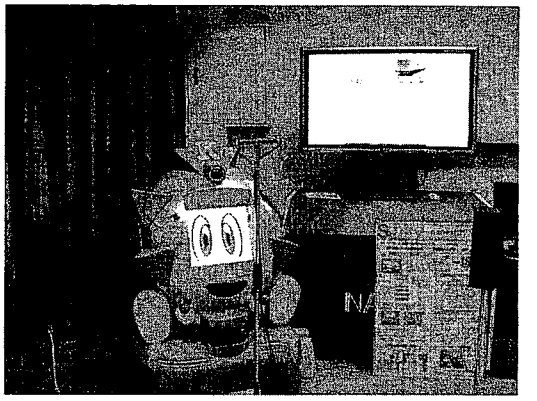

図 2 ハンズフリー音声対話ロボット "キタロボ"
Fig.2 Hands-free Spoken Dialogue Robot "Kitarobo".

前処理であるユーザの発話区間の特定，音声区間検出の性能は高く，認識性能も高い精度を得られる．しかしながら従来のインターフェースにおいて高い認識精度を得るためには，ユーザがマイクを装着したり，マイクに接近していなければならないといった制約がある．カーナビゲーションの音声認識機能を利用するためにマイクを持って話すことは危険であり，またヘッドセットを装着することは煩わしい．音声案内システムやロボットとの対話においても，子供であればシステムに充分に近付いて話しかけるが，大人は羞恥心等の理由からか一歩二歩引いた位置から話す状況が多々見られる．つまり多くの音声認識システムでは人と人とが対話するようにユーザに負担の無い自然な発話や対話を行うことを可能にするハンズフリー音声認識の導入が求められる．

また本研究室では実環境で「たけまるくん (屋内，公共施設)」[1] や「キタちゃん，キタロボ (屋外，駅構内)」[2](図 1) といった音声対話システムを運用しており，様々な音声関連の研究に用いている．これらは指向性マイクを使用しているが，現在マイクロホンアレーを用いたハンズフリーインターフェースへの置き換えにより実環境ハンズフリー音声対話システムの実現に向け研究を進めている．図 2 は実験室にて実装したシステムの外観図である．我々はシステムの周囲に配置された最大 32 個あるスピーカーを用いて，実際に駅や博物館で収録した環境音を音場再現することで実環境を模擬しながら実験を行っている．

ハンズフリー音声認識システムの問題点として，ユーザとマイクの位置が離れることによりマイクに入力されるユーザの直接音のパワーが減衰したり，また実環境においては様々な背景雑音が混入するために入力音声の SNR は低下するといった問題がある．SNR が低下するにつれて発話と背景雑音の特徴差が曖昧になり音声認識の性能は低下する．特に音声区間検出の性能が著しく低下する．音声区間検出は入力音声からユーザの発話した区間を特定する処理で，通常音声認識の前に行われる．もし正しく音声区間の検出ができなければ，システムはユーザの発話した区間とは異なる区間を認識することになる．つまり音声区間検出の失敗はすなわち音声認識の失敗に繋がる．従来の音声区間検出では振幅レベルや零交差数情報を利用したパワーに基づく手法 [3] [4] が利用されているが，この手法の性能は SNR と密接に関係があり，SNR の低下に伴い検出性能も大きく劣化し，SNR が 0dB に近付くにつれ利用が困難となる．入力機器の指向特性を高めたり，入力信号に音響信号処理を施すなどして SNR を改善するなどの対策がよく講じられるが，実環境ハンズフリー音声認識においてはパワーに基づく音声区間検出手法では現状十分な性能とは言えない．他の音声区間検出手法として，ユーザ自身が発話を任意に収録する "Push and Talk" 手法は雑音環境における音声区間検出としては効果的ではあるが，発話の際に手が拘束されてしまうなど容易に使えなかったり対話システムに向かないといった問題がある．他に自己相関関数のピーク値 [5] や，雑音抑圧後のスペクトルの分散値 [6] 利用した手法，エントロピー [7] や，最小平均二乗誤差推定した音声スペクトルから求められる対数尤度を使った確率モデルに基づく手法 [8]，フレームベースの Gaussian Mixture Model (GMM) に基づく音声区間検出 [9] などが報告されているが，実環境ハンズフリー音声認識における自然発話のリアルタイム認識となるとまだまだ効果は不充分である．

我々は実環境ハンズフリー音声認識において頑健な音声区間検出および認識を実現するために音響モデルと言語モデルを利用した認識に基づく音声区間検出および認識アルゴリズムを提案した [10]．本稿では，音響モデルと言語モデルに基づく音声区間検出を用いたハンズフリー音声認識アルゴリズムを評価するために既存の音声区間検出手法を用いた場合とで区間検出や認識性能について比較・評価を行う．既存音声区間検出手法として，振幅レベルおよび零交差数に基づく手法 [3]，Sohn らが提案した確率モデルに基づく手法 [8]，また GMM に基づく手法 [9] を挙げる．

2. 従来の音声区間検出手法

2.1 振幅レベルおよび零交差数に基づく音声区間検出

従来使用される音声区間検出手法の一つにパワーに基づく音声区間検出 [3] が挙げられる．この手法は入力信号のエネルギーに対して閾値を設け，閾値を越えた区間を発話区間として検出する手法である．だが発話の始端や終端が子音であるときなどはパワーが小さく，正確に区間検出することは難しい．そこで閾値によって検出された区間の前後数フレームを発話区間に含めることで文頭や文末が切れてしまうことを防ぐ．また，発話の始端や終端に存在すると考えられる摩擦音を零交差回数に対する閾値で検出する方法がある．これらの手法は入力信号のエネルギーを基準つまり SNR の差を利用しているため，マイクに接話した時のような SNR の高い音声であれば性能が高いが，ハンズフリー音声認識のような背景雑音と目的音の差が小さい場合，雑音の区間を発話区間と誤って検出されてしまうといった問題がある．

2.2 確率モデルに基づく音声区間検出

雑音環境下においても頑健に音声区間検出を行う手法として，Sohn らにより提案された確率モデルに基づく手法 [8] がある．この手法は観測される信号が音声状態と背景雑音状態を遷移する信号であると仮定して，観測信号が各状態に属する確率つまり尤度の比を求めて，閾値により音声か背景雑音かの判定を行う．また尤度比は単に各フレーム単位で計算するのではなく過去のフレームの状態を考慮して計算される．この手法は音声と背景雑音の遷移モデルを定義はしているが，実際の音声で学習されたモデルを用いているのでは無く，最小平均二乗誤差推定により求めた音声スペクトル [11] 情報を用いる．そのため，この手法の性能は SNR の推定精度に強く依存してしまう問題がある．また前提として，雑音が既知で定常的であることを条件にしており，実環境に存在する様々な非定常雑音に対応することが難しい．

2.3 GMM に基づく音声区間検出

SNR に依存せず，音声と雑音の音響的特徴の違いを利用して音声区間検出を行う手法として混合正規分布モデル (GMM) に基づく手法 [9] がある．この手法は音声と背景雑音の音響的特徴をあらかじめモデル化しておき，入力の各フレームが音声か背景雑音かどちらの尤度が高いかによって判定を行う．この手法は，音声と雑音の音響的特徴のみを用いて発話区間を検出するので，SNR が比較的悪い条件でも音声と背景雑音を区別できるが，背景雑音が音声 (背景会話) である場合や，モデルの学習が不十分であったりシステムを使用する環境に適応されていない場合は，目的音声と背景雑音を誤って検出してしまうといった問題がある．

3. 音響モデルと言語モデルを用いた認識に基づく音声区間検出および認識アルゴリズム

3.1 音響モデルと言語モデルを用いた認識に基づく音声区間検出

我々は以前，SNR に依存せず，また音響的特徴だけでなく言語情報を用いて音声区間検出を行う手法 [10] を提案した．図 3 は提案した音声区間検出の概要図を示す．提案手法では従来の音声区間検出のように認識の前処理として働くのではなく，認識処理と一体化して，常に認識計算を行う中で音声区間検出と認識仮説導出を同時に行う．具体的には音声区間検出は認識結果に基づいて行われ，そのアルゴリズムとして，まずシステム起動時やユーザが発声していない区間のように，認識仮説候補の第一位が無音単語 (背景雑音) である場合は検出は開始されず，ユーザが発声して任意の単語 (発話) が仮説候補の第一位に上がった時，区間検出は開始される．区間検出開始後，再び認識仮説候補の第一位が無音単語になり，その状態がしばらく (ここでは 30 から 35 フレーム程度) 継続 (ユーザの発声が終了) した場合に区間検出を終了し，以上の区間を音声区間として検出する．通常の音声認識では認識計算が行われる前に音声区間検出が行われ，検出された区間のみを認識対象とするが，提案手法では音声区間検出のために認識結果を用いるため全ての入力が認識対象，つまり常時認識を行うことになる．提案した認識による音声区間検出はフレーム同期ビーム探索により達成される．提案手法ではフレーム同期ビーム探索を用いることにより，音声区間検出を行うだけでなく同時に検出した区間の認識仮説も求めている．よって提案手法は音声区間検出と認識計算を同時に行うアルゴリズムとなっている．なお図 3 中で音声区間検出トリガがかかった時，探索が巻き戻って再探索が行われるが，フレーム同期ビーム探索は高速で探索を行うため，区間検出が終了する頃にはリアルタイムに計算が追い付くためディレイは発生しない．また探索の再開位置を検出トリガ地点よりもいくらか (およそ 300 から 350msec 程度) 前に設定する理由は，音声認識においてユーザの発話の前に不必要に長い無音が入っていると認識性能が劣化するため，必要なだけの無音区間をユーザの発話の前に付与するためと，提案手法は常時認識を行うため，探索位置の再設定処理を行わない場合ほとんどユーザの発話とは関係の無い区間が検出対象になってしまうためである．また，一つの音声区間検出が終了し探索が再開する際，区間検出終了後すぐにユーザの発話が入力される場合に対処するために探索を区間終了位置以前から再開する．

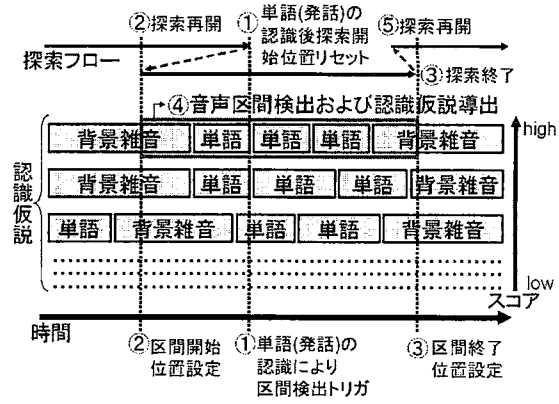

図 3 提案した音声区間検出概要図
Fig.3 Outline of Proposed Voice Activity Detection.

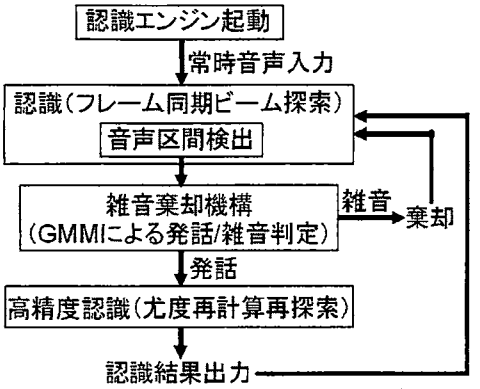

図 4　提案音声区間検出手法を用いた認識アルゴリズム
Fig.4 Recognition Algorithm using proposed Voice Activity Detection Method.

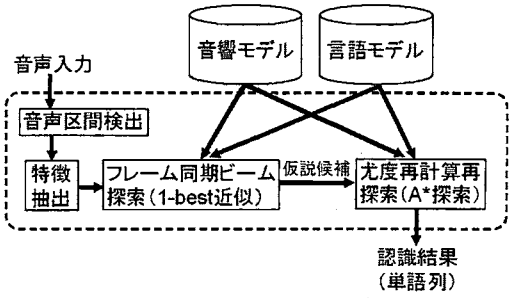

図 5　音声認識エンジン *Julius* の構成図
Fig.5 Outline of System Organization of Julius.

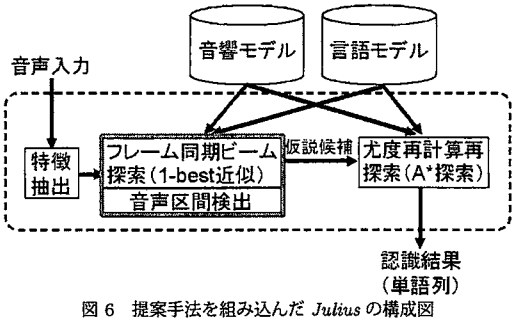

図 6　提案手法を組み込んだ *Julius* の構成図
Fig.6 Julius that Integrated the Proposed Method.

提案手法の音声区間検出性能はモデルの性能，特に音響モデルの性能に強く依存することが先行研究 [12] により確認されている．雑音環境下で正確に音声区間検出を行うためには音声は音声，背景雑音は背景雑音と正しく認識する必要がある．そのためシステムを運用する環境に合わせて音響モデルを適応することで効果的に提案手法を利用することが可能となる．

3.2　提案音声区間検出手法を用いた認識アルゴリズム

提案した音声区間検出手法はフレーム同期ビーム探索により認識結果も同時に求める．だがフレーム同期ビーム探索は高速ではあるが認識精度は充分に高いとは言えないため，結果としてこの探索だけでは認識性能は充分では無い．また，実環境では音響モデルが充分に学習されている場合でも学習外の未知の雑音が入力される可能性があり，その場合，ユーザの発話でなくても探索結果として認識仮説候補の第一位が無音単語以外の任意の単語になりうる．そして提案手法ではそのような場合でも音声区間検出が行われてしまうため，常時認識を行う中で不要な結果を求めてしまう可能性があると予想される．

そこで我々は提案手法を用いたハンズフリー音声認識認識アルゴリズムとして，2 つの機構を採用する．まず 1 つ目に，認識性能の改善のために音声区間検出時に得られた仮説候補を用いて尤度再計算再探索を行うことで高精度の認識結果を求める．2 つ目に音響モデルでは学習しきれない雑音，例えば咳や笑い声，物音や風の音などをあらかじめ学習した GMM を用いることで不要な区間を棄却する．提案アルゴリズムの認識の流れを図 4 に示す．これらの機構を取り入れ，我々は提案した認識アルゴリズムを音声認識デコーダ Julius [13] [14] に実装した．

3.3　Julius 概要と提案アルゴリズムを導入した Julius

Julius は音声認識の研究や音声認識システムの開発に利用される高性能かつリアルタイムな汎用大語彙連続音声認識エンジンである．Julius はライセンスフリーかつオープンソースソフトウェアで，Linux や Windows をはじめ様々なプラットホームで動作する．また高い汎用性を持っており，音響モデルや言語モデル，辞書など音声認識の各モジュールを入れ換えることで多言語に対応可能であるなど，幅広い用途に応用できる．

現行の Julius(ver3.5.3) の構成の概要を図 5 に示す．また提案アルゴリズムを組み込んだ Julius の概要を図 6 に示す．現行の Julius では通常，音声入力があると振幅レベルおよび零交差数に基づき音声区間検出を行う．次に MFCC 特徴量を抽出し，それから認識計算 (探索) を行う．認識計算は HMM 音響モデルや統計的言語モデルなどに基づき行われる．そして探索には 2 パス探索アルゴリズムが用いられる．まず 1 パス目にフレーム同期ビーム探索を行い認識仮説を絞り込む．この時 GMM 計算も同時に行われ，入力が発話であるか雑音であるかを判断する．もし入力が発話と判断された場合，次に 2 パス探索として尤度再計算再探索が行われ，認識結果を求める．入力が雑音である場合，2 パス探索は行われず，入力は棄却される．

提案アルゴリズムを導入した Julius では．まず常時入力音声を特徴抽出し，フレーム同期ビーム探索により音響モデルと言語モデルに基づき音声区間検出および認識仮説候補の導出を同時に行う．次に検出した区間を GMM に基づき発話/雑音判定を行う．もし検出した区間が発話であれば尤度再計算再探索により認識結果を求め，雑音であれば検出し区間を棄却し，音声区間検出のための探索を再開する．

4.　評価実験

本研究は音声案内システム「たけまるくん」や「キタちゃん・キタロボ」などに適用されるハンズフリー音声認識アルゴリズムの構築と実現を目標としている．そのため主に対話システムに必要な音声認識性能に関して従来手法と提案手法を比較評価する．また音声区間検出性能の考察も行う．

4.1　実験条件

従来手法の評価のために図 5 で構成される Julius を用いた．また提案手法の評価は図 6 で構成した Julius を用いた．評価データとして，駅構内に設置したキタちゃん (図 1，屋外環境) を用いて，話者とマイクの距離を変えて (接話，1m，1.5m) 収

表 1 実験条件
Table.1 Experimental Condition

入力音声	条件 1	接話, SNR≒50dB
(2 と 3 はハンズ	条件 2	約 1m 離れた位置, SNR≒10dB
フリーに相当)	条件 3	約 1.5m 離れた位置, SNR≒6dB
音響モデル		2000 状態, PTM, Gaussian
音響特徴		12MFCC, 12ΔMFCC, ΔE
音響モデル学習		Baum-Welch, 3 Iterations
音響モデル適応		MLLR-MAP, 3 Iterations, 256 Classes
言語モデル		3-gram, Kneasr-Ney smoothing
		Vocabulary size is 40k.
発話タスク		情報案内 (駅構内, 路線, 観光, 施設, 周辺, 天気, News), 挨拶や対話システム自身に対する質問など
評価データ		話者 1 名, 204 発話, 1024 単語, 未知語率 0 %

録した SNR の異なる 3 つの音声を用いる. SNR はそれぞれ, 接話は約 50dB, 1m 離れた位置からの音声は約 10dB, 1.5m の位置からの音声は約 6dB 程である. 各評価データはどれも同じ文章からなる全 204 発話であり, それぞれ発話と発話の間で収録を区切らずに大体 3〜4 秒程の非発話区間を含めた 204 発話で 1 つの音声データである.

使用する音響モデルは JNAS を初期モデルとし, たけまるくん (屋内環境) で収録した過去 2 年間分のデータ (大人 23,417 発話, 子供 120,671 発話) を用いて学習・構築し, キタちゃんで収録された一ヶ月半のデータ (大人 6,661 発話, 子供 9,472 発話) を用いて駅環境に MLLR 適応を行った環境適応モデルを使用する.

従来手法では, 各手法における閾値設定やモデル学習の精度次第で音声区間検出および認識性能が変化する. そのため本稿では, まず振幅レベルと零交差数に基づく手法では振幅レベルや零交差数の閾値を様々に変化させて実験し, 最も精度の高かった結果を採用する. 確率モデルに基づく手法においても, 対数尤度比検定のための閾値や分析フレーム長等のパラメータを変えて実験し, 最も良好であった結果を示す. GMM に基づく手法ではあらかじめ評価データと同じ環境 (キタちゃん) で収録した約半年間の音声データ (15,091 発話) を用いてモデルを学習・構築し, 実験を行った. 提案手法および従来手法ともに発話の前後区間に付与する無音区間長等は全て同条件に設定した.

またケプストラム平均正規化 (CMN) 処理を行っており, 前回の有効な入力の最後部 5 秒分のケプストラム平均を用いている.

4.2 実験結果

まずはじめに予備実験として, 先行研究 [12] で示した提案手法の音響モデルの環境適応有無による認識性能の違いを図 7 に示す. 予備実験で使用した音響モデルは, 節 4.1 で記述した環境適応モデルと異なる実環境のモデルとしてたけまるくん (屋内環境) モデルおよびクリーン環境として JNAS モデルである. 結果から提案手法は音響モデルを環境適応することで性能が改善されることが分かる.

次に提案手法と従来手法の認識実験の結果を図 8 に示す. 結果から接話入力の場合のような SNR が高い場合は提案手法,

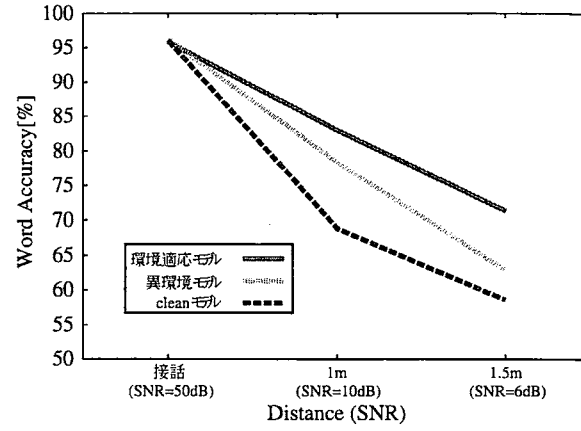

図 7 提案手法の音響モデルの環境適応有無による結果
Fig.7 Result of Proposed method using Acoustic Modelwith or without adaptation to environment.

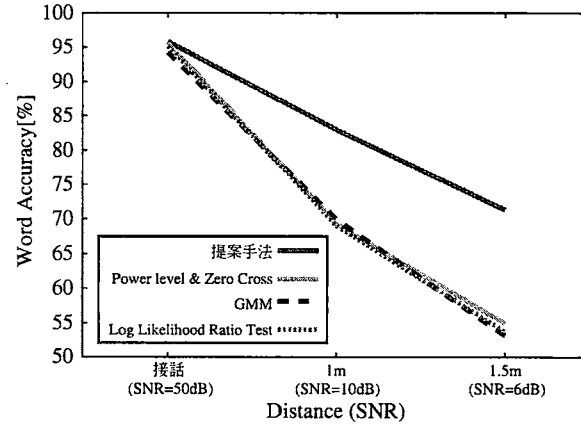

図 8 実験結果 (単語正解精度)
Fig.8 Experimental Result (Word Accuracy).

従来手法ともに高い性能を示しているが, ハンズフリー音声入力のように SNR が低い場合, 提案手法と従来手法の性能差が際立ち, 提案手法の性能が高いことが確認できる.

音声区間検出性能について検討を行うため表 2 から 4 に認識誤りと VAD 誤りの結果を示す. 表中の False Acceptance (FA) は本来発話区間では無い区間を発話として検出した回数. つまり評価対象である全 204 発話以外の区間を誤って検出した回数を表しており. False Rejection(FR) は本来発話区間である区間を検出できなかった回数. つまり全 204 発話のうち何発話検出できなかったかを表している. なおこれらの結果は認識時の GMM 判定処理を経て求められた結果のため, 音声区間検出されていたとしても, 検出区間があまりに短かったり人の声であっても検出区間によっては雑音として棄却された可能性がある. 結果から提案手法は従来手法よりも検出性能が高いことが分かる. 提案手法は FR が 11 発話あるにもかかわらず, ほとんど FR の無い従来手法よりも削除誤り率が低い. これは従

来手法では文頭や文末が検出され難く，ほとんど全ての音声に対して文頭落ちあるいは文末落ちが発生しているのに対し，提案手法ではそれらが発生していないからだと考えられる．従来手法の問題点としてパワーに基づく手法は接話であっても文頭や文末の検出が難しい上にハンズフリー音声認識では発話中であっても背景雑音に音声が埋もれてしまうためにさらに性能が低下してしまう．確率モデルに基づく手法では雑音の逐次推定を行っていないため様々な雑音の存在する実環境では推定が困難となり性能が低下すると考えられる．GMM に基づく手法でも，ハンズフリー音声認識では文頭や文末が完全に背景雑音に埋もれてしまうことが多く，当該区間は背景雑音フレームとして判定されやすくなってしまう．そのため振幅パワーに基づく手法と似た結果になったと考えられる．

GMM による音声区間検出手法では音声と背景雑音を判別する GMM を作成する際，キタちゃんで収録した半年分のデータを使用しており，音響モデルの環境適応時に使用したデータの一ヶ月半のデータよりも大量のデータを用いてモデルを構築しているため提案手法以上に音声・背景雑音の判別能力が高いはずだが，検出や認識性能は提案手法の方が優れている．このことから単フレーム，あるいは数フレームの音響特徴だけで音声区間検出を行うよりも言語情報を用いてユーザの発話を検出することが有効であると考えられる．

5. ま と め

本稿では音響モデルと言語モデルに基づく音声区間検出を用いたハンズフリー音声認識アルゴリズムの評価を行うため，従来の音声区間検出手法との比較実験を行った．実験の結果，接話のように入力の SNR が高い場合は提案手法，従来手法ともに性能が高い．だがハンズフリー音声認識のように SNR が低い場合は従来手法，提案手法ともに性能が下がるものの，提案手法は従来手法よりも音声認識性能，音声区間検出性能ともに大幅に改善されることを示した．実環境ハンズフリー音声対話システムへの導入を考慮した際，提案手法は雑音に対しては頑健であるが，背景会話のような人の声が背景雑音である場合性能が劣化してしまう．そこで今後はハンズフリー音声対話システムの性能向上のために音源方向推定などの信号処理と組み合わせたシステムの構築を行う予定である．また使用する音響モデルに関して，モデルを作成する際の学習データは接話収録された音声を使用したが，ハンズフリー収録された音声を用いて学習を行った際，性能が改善されるか評価する必要があると考えられる．

6. 謝　　辞

本研究は文部科学省のリーディングプロジェクト [e-Society 基盤ソフトウェアの総合開発] により実施したものである．

表 2　条件 1(SNR≒50dB) における認識誤りと VAD 誤り
Table.2 Recognition and VAD Error in Condition 1(SNR≒50dB)

手法	認識誤り率	削除誤り率	挿入誤り率	FA	FR
提案手法	2.45%	0.69%	0.98%	3	0
振幅パワー	2.65%	0.98%	0.69%	0	0
確率モデル	2.65%	1.37%	0.78%	15	0
GMM	2.55%	1.67%	1.47%	7	1

表 3　条件 2(SNR≒10dB) における認識誤りと VAD 誤り
Table.3 Recognition and VAD Error in Condition 2(SNR≒10dB)

手法	認識誤り率	削除誤り率	挿入誤り率	FA	FR
提案手法	9.50%	6.95%	0.59%	0	3
振幅パワー	15,21%	14.62%	0.88%	1	0
確率モデル	14.80%	15.00%	1.18%	15	0
GMM	13.86%	13.57%	2.65%	4	1

表 4　条件 3(SNR≒6dB) における認識誤りと VAD 誤り
Table.4 Recognition and VAD Error in Condition 3(SNR≒6dB)

手法	認識誤り率	削除誤り率	挿入誤り率	FA	FR
提案手法	12.65%	15.29%	0.69%	0	11
振幅パワー	17.96%	26.50%	0.59%	5	0
確率モデル	17.78%	27.41%	1.08%	24	0
GMM	17.29%	27.37%	2.17%	7	5

文　　献

[1] R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano, "Public Speech-oriented Guidance System with Adult and Child Discrimination Capability, " proc.of ICASSP, pp.433–436, 2004.

[2] H. Kawanami, M. Kida, N. Hayakawa, T. Cincarek, T. Kitamura, T. Kato and K. SHikano, "Spoken Guidance Systems Kita-chan and Kita-chan robot.Their Development and Operation in a Railway Station, " tech.rep., IEICE, SP2006–14, 2006.

[3] L.R.Rabiner and M.R.Sambur, "An algorithm for determining the end-points of isolated utterances, " The Bell System Technical Journal, Vol.54, No.2, pp.297-315, 1975.

[4] ITU-T Recommendation G.729 Annex B., 1996.

[5] Kristjansson Trausti, Deligne Sabine, Olsen Peder, "Voicing features for Robust speech detection, " Interpeech 2005, pp.369–372, 2005.

[6] ETSI ES 202 050 Recommendation, "Speech Processing, Transmission and Quality aspects(STQ);distributed speech recognition; Front-end Feature Extraction Algorithm; Compression Algorithms, " 2002.

[7] Jia-lin Shen, Jeih-weih Hung, Lin-shan Lee, "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environment, " Proc. International Conference on Spoken Language Processing (ICSLP), 1998.

[8] Jongseo Sohn, et al. "A statistical Model Based Voice Activity Detection, " IEEE Signal Processing Letters, Vol.6, No.1, 1999.

[9] Norbert Binder, Konstantin Markov, Rainer Gruhn, Satoshi Nakamura, "Speech/Non-Speech Separation with GMMs, " Proc.of ASJ Fall Meeting, Vol1, pp.141–142, 2001

[10] 酒井他，"実環境ハンズフリー音声認識のための音響モデルと言語モデルに基づく音声区間検出と認識アルゴリズム，" 信学技報，SP2007-17，pp.55–60，2007.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, " IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-32, pp.1109–1121, 1984.

[12] 酒井他，"実環境ハンズフリー音声認識のための音響モデルと言語モデルに基いた音声区間検出の評価，" 音響学会，2007 年秋季研究発表会，3-3-12，pp.167–168，2007.

[13] Open-Source Large Vocabulary CSR Engine Julius developed by A.LEE
introductory web pages at: `http://julius.sourceforge.jp/`

[14] A. Lee, T. Kawahara and K. Shikano. "Julius - an open source real-time large vocabulary recognition engine", Proc Eurospeech2001, pp1691–1694, 2001