『経営学論集』，第18巻　第２号，67-85，2007
KYUSHU SANGYO UNIVERSITY, *KEIEIGAKU RONSHU*（*BUSINESS REVIEW*）Vol. 18, NO. 2, 67-85, 2007

〔論　説〕

有声音の調波構造を利用した雑音に頑健な音声区間検出手法

菅　野　禎　盛

九州産業大学経営学部

sugano@ip.kyusan-u.ac.jp

［概　要］

㈶九州システム情報技術研究所にて開発された騒音下音声認識システムを用いて，入力信号に音声信号が含まれる区間の検出を行う処理を作成し，その性能に関する定量的評価を行った。この騒音下音声認識システムは，入力された音声信号に含まれる調波構造を検出し，音声特徴ベクトルを推定するものである。音声区間の検出のためには，音声特徴ベクトルの推定よりも音声の調波構造，および基本周波数の推定を高精度で行う必要がある。そこで，本システムではこの騒音下音声認識システムを基本周波数の推定用に特化させるように修正し，さらに音声区間の検出処理を作成した。提案手法の検出性能の評価を複数の男性話者による発話音声を用いて行った結果，クリーン音声に対しては従来法と比較して約11％の検出性能の向上が示され，また，騒音が重畳された音声に対しては，従来法と比較して，SNR（signal to noise ratio）が10dBの条件と０dBの条件でそれぞれ約３％の検出性能の向上が示された。さらに本システムは従来法と比較してより安定して音声区間の検出を行うことができることも示された。

キーワード：自動音声認識，音声区間検出，非定常騒音，調波構造

１　はじめに

コンピュータによる自動音声認識技術は，いわゆる「人に優しい」ヒューマンマシンインターフェースを実現するための核となる技術のひとつである。その成果は，パソコンへの音声入力，カーナビゲーションシステムへのハンズフリーな音声入力，電話を介した商品注文システム（テレフォニーシステム），音声による個人認証など，社会の様々な場面で利用されている。

自動音声認識では，音声が存在する区間を正確に検出することが極めて重要である。発話が静かな環境ではっきりとなされている場合は，信号レベルに適当な閾値を設けることにより比較的容易に音声区間とそれ以外の雑音区間を切り分けることができる。しかし，自動音声認識

システムが騒音のない理想的な環境で用いられることはあまりなく，もっぱら騒音下で利用されることが多い。例えば，車内騒音下でのカーナビゲーションシステムへの音声入力，環境騒音下での自動券売機への音声入力，などである。このように騒音下で発話がなされる場合，周囲の雑音や他の人の話し声などが混入してしまうために，静かな環境での発話と比較して音声区間を正確に検出することが非常に困難となる［12］。そして音声区間の誤検出は即座に，その後のパターンマッチング処理の性能を悪化させる（認識率の低下）。そのため，雑音に対して頑健な（雑音の有無に左右されない）音声区間検出法の開発が強く望まれている。

本研究では，㈶九州システム情報技術研究所で開発された自動音声認識システムであるPHONOBEST（PHOnetic kNOwledge Biased ESTimation）を利用した雑音に頑健な音声区間検出手法の開発と性能評価を行った。PHONOBESTは音声の基本周波数（F_0）に関する情報を積極的に利用して音声認識を行うシステムである。これまでの研究では，従来法と比べてPHONOBESTが非定常で予測不可能な騒音下でも比較的よく音声（単語）を認識できることが示されている［8，9，10，11］。

従来の音声区間検出手法は多かれ少なかれ，音声に重畳されている雑音の特徴が一定時間内では変化しない（定常騒音）という仮定を設けている。例えば，スペクトルサブトラクションという方法は，まず，雑音しか存在しないとはっきり分かっている信号部分をスペクトル分析し，雑音のスペクトルを得ておく。音声と雑音を分離する際には，入力された信号のスペクトルから，予め計算しておいた雑音のスペクトルを引くことで音声のみのスペクトルを得る［2］。これを時間領域に戻すことで雑音が除去された信号が得られ，容易に音声区間を検出することができる。しかしこの方法では，雑音が定常でない場合は得られる音声信号が歪み，結果的に音声区間の誤検出が生じる。

また，信号処理の段階で音声を雑音から分離するのではなく，パターンマッチング処理の段階で音声区間を検出する方法もある。この方法ではまず，音声と雑音のそれぞれを認識できるHMM（Hidden Markov Model）を用意し，それぞれのHMMは音声と雑音によって学習しておく。そして音声HMMと雑音HMMを連結したHMMを使って，雑音が重畳した音声信号を分析する［13］。その結果，音声認識と音声区間の検出を同時に行うことができる。しかしこの方法も，予め学習した雑音とは異なる雑音が音声に重畳している場合には対応が難しく，誤認識・誤検出が生じてしまう。

このように，いずれの方法も雑音の特性について予め何らかの仮定を設ける必要があるという問題点を抱えている。つまり，音声認識システムを運用する環境にどのような雑音が存在するかが事前に分かっていなければならない。このような制約は自動音声認識システムの汎用性

を大きく狭めることになる。

PHONOBESTは音声に重畳される雑音の性質，および音声と雑音との信号比（S/N比）について何らの事前仮定を設ける必要がないという特徴を持っている。この特徴は，様々な種類の騒音下で音声認識を行う必要があるというシステム運用上の要請からすると，非常に有利な特徴である。また，PHONOBESTは現在主流となっている音声認識方式との整合性がよい。現在主流のほとんどの音声認識システムは，HMMのような確率統計的な枠組で実現されている。PHONOBESTはHMMを用いた既存の音声認識システムに容易に組み込むことができる。

本研究は，騒音下での自動音声認識システムとして開発されたPHONOBESTを騒音下での音声区間の検出に利用すべく拡張することを目的とした。まず，PHONOBESTの概略について述べ，次にPHONOBESTを音声区間検出に利用するための拡張方法について述べる。最後に，開発した音声区間検出システムと従来法の性能評価実験について述べ，提案手法の長所と短所についてまとめる。

2　PHONOBESTの概要と音声区間検出に向けた拡張について

2.1　PHONOBESTの概要

PHONOBESTは，音楽信号のメロディとベースラインを推定する手法であるPreFEst (Predominant-F0 Estimation Method) [3, 5] を，離散HMMに基づく音声認識システムの一部として組み込むことができるように拡張したシステムである。

図1にPHONOBESTが実行する処理の概念図を示す。PHONOBESTが実行する処理は，入力信号のスペクトル分析処理，複数の音モデルの生成処理，そして入力信号中に各モデルがどの程度含まれているかを推定する処理，という3つの大きな処理に分けることができる。

PHONOBESTでは，まず観測信号を周波数分析して観測信号の詳細な周波数スペクトルを得る。そして，得られた周波数スペクトルを確率密度関数 $p_{obs}^{(t)}(x)$ と見なし，$p_{obs}^{(t)}(x)$ が事前に作成した音モデル $p(x \mid F, m, c(h \mid F, m))$ の重み付き混合分布から生成されたと見なす。

$$p_{obs}^{(t)}(x \mid \theta^{(t)}) = \int_{Fl}^{Fh} \sum_{m=1}^{M} w^{(t)}(F, m)\, p(x \mid F, m, \mu^{(t)}(F, m))\, dF \tag{1}$$

$$\mu^{(t)}(F, m) = \{c^{(t)}(h \mid F, m) \mid h = 1, \cdots, H\} \tag{2}$$

ここで，F は，想定される基本周波数の候補，m はベクトル量子化コードブックのセントロイドを，M はその個数を表す。H は考慮する倍音の数である。左辺の $p_{obs}^{(t)}(x \mid \theta^{(t)})$ は観測信号を周波数分析した結果得られたスペクトルを確率密度関数化したものである。右辺の

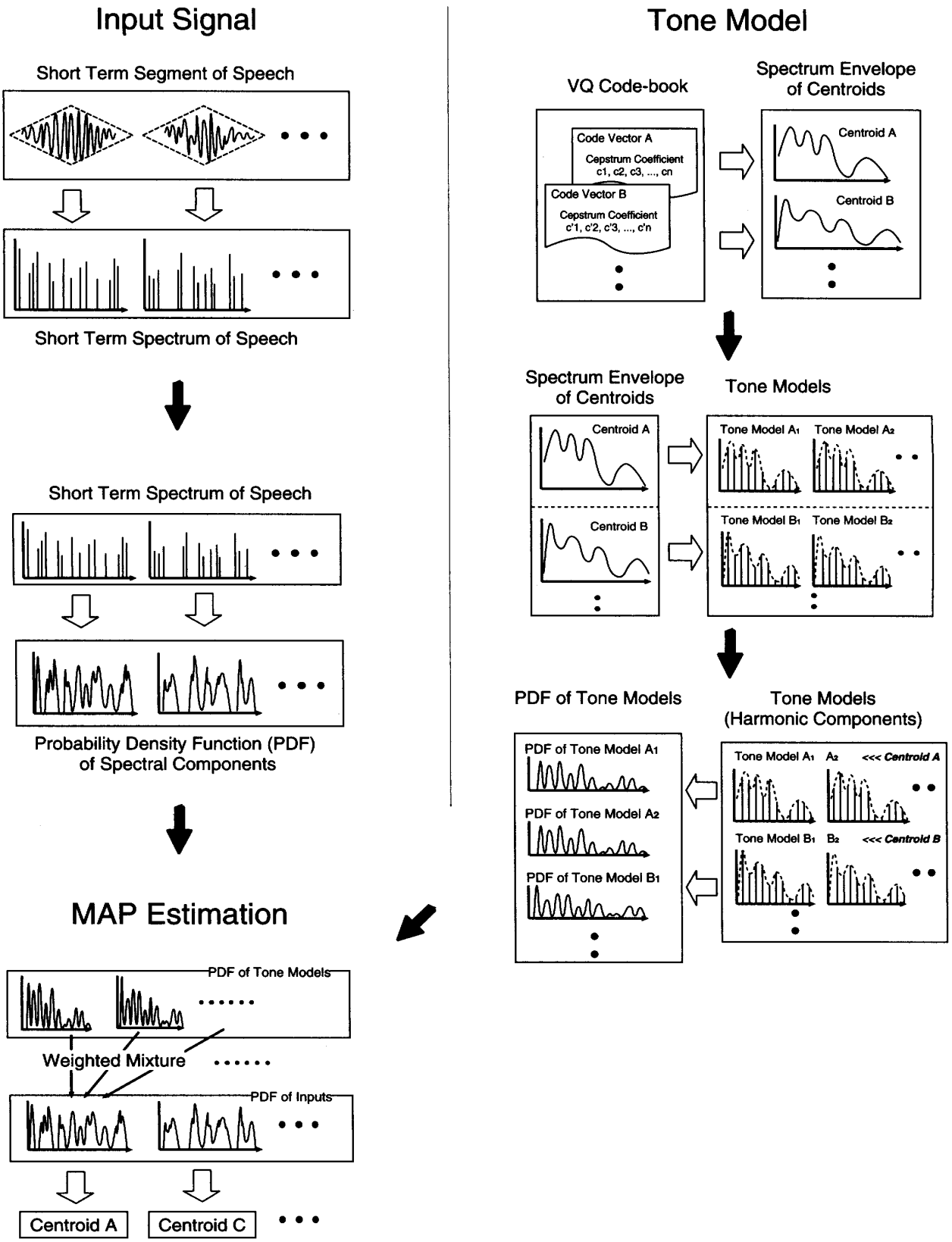


図1　PHONOBESTが実行する処理の流れ

$p(x \mid F, m, \mu^{(t)}(F,m))$は，基本周波数が$F$であるセントロイド$m$の音モデルの確率密度関数であり，$w^{(t)}(F,m)$はその音モデルに対する相対的重みづけを表すパラメータである。式(1)は，基本周波数Fについての積分形となっているが，実際の処理ではある一定の周波数間隔でサンプリングした周波数値について和をとる形となる。

また，基本周波数がFであるセントロイドmの音モデルの確率密度関数は，

$$p(x \mid F, m, \mu^{(t)}(F,m)) = \sum_{h=1}^{H} p(x,h \mid F, m, \mu^{(t)}(F,m)) \tag{3}$$

$$p(x,h \mid F, m, \mu^{(t)}(F,m)) = c^{(t)}(h \mid F,m)\, G(x \mid F,h) \tag{4}$$

となる，ここで，Hは考慮する倍音の数，Gは周波数$F \cdot h$で最大値を持つガウス分布である。また，$c^{(t)}(h \mid F,m)$は基本周波数Fの音モデルmが持つ倍音成分の相対的振幅を表すパラメータである。

PHONOBESTは，パラメータ$c^{(t)}(h \mid F,m)$と$w^{(t)}(F,m)$に関する事前分布を用いて，観測信号が与えられたときのもっとも起こりやすいパラメータ$c^{(t)}(h \mid F,m)$と$w^{(t)}(F,m)$をEMアルゴリズムを用いた最大事後確率推定（MAP推定）により推定する。これらのパラメータの事前分布は何らかの既知の情報に基づいて予め適当に設定しておくか，無情報一様事前分布を用いる。PHONOBESTのパラメータ推定方法の詳細については，文献［8，9，10，11］を参照されたい。

PHONOBESTのポイントは，有声音が基本周波数とその倍音からなる調波構造を持つという性質を積極的に利用し，倍音部分のみの情報を入力音声とテンプレート(音モデル)のパターンマッチングに利用するという点である。従来の音声認識システムの多くが，スペクトル包絡またはケプストラム係数のような全周波数領域に関する情報をテンプレートとして持つのと比べると，調波構造部分の周波数情報を積極的に利用するという点がPHONOBESTのユニークな点である。

2.2 PHONOBESTの音声区間検出への拡張

前述のように，PHONOBESTはまず候補となる基本周波数のすべてについて倍音構造を持つ音モデルを用意し，入力信号をこれらの音モデルの重み付け和により表現する。そして次に，EMアルゴリズムにより入力された音声信号の各分析フレームごとに各音モデルの重み値を計算し，最も重み値が高くなった音モデルをそのフレームでの推定結果とする。

PHONOBESTが入力音声信号の各分析フレームで推定する各音モデルの重み値$w(F,m)$は，基本周波数の候補Fと，ベクトル量子化コードブックのセントロイドmの関数となっている。

ここで，重み値$w(F,m)$の全てのセントロイドmに関する和$\sum_{m=1}^{M} w(F,m)$は，基本周波数Fの関数となっており，もしその分析フレームでの入力音声が強い倍音構造を持っている場合，すなわち有声音である可能性が高い場合は$\sum_{m=1}^{M} w(F,m)$の値がある特定の基本周波数候補に関して非常に大きい値をとり，その他の基本周波数候補については非常に小さい値をとる。そのため$\sum_{m=1}^{M} w(F,m)$の推定値は有声／無声の検出に利用することができる。この考え方が提案手法のポイントである。

しかしながら，音声は有声音だけから構成されているわけではない。当然のごとく無声音も音声には含まれる。そこで本研究で提案する音声区間検出手法では，まず$\sum_{m=1}^{M} w(F,m)$の値により入力音声信号の有声区間を検出し，その結果を利用して音声区間を検出する，という2段階の処理を行う。

また，PHONOBESTは元々音声認識システムとして開発されたため，基本周波数Fの推定よりも音モデルmの推定に重点をおいて設計されている。これは具体的には，パラメータ$w(F,m)$の推定において，基本周波数Fの候補の数を抑え，音モデル（コードブックのセントロイド）mの候補の数を多くしてあるということである。しかし，有声／無声の判定を行うためにはセントロイドの推定ではなく基本周波数の推定に重点をおくようにPHONOBESTを修正する必要がある。そこでまず，基本周波数の候補を80～240 Hzの周波数範囲内で2 Hzおきに細かく設定し，高い精度で基本周波数が推定できるようにPHONOBESTの設定を修正した。

しかし，このように基本周波数の高精度の推定を行おうとすると，計算量が膨大になり，高速な実行ができなくなる。そこでまず，入力信号の周波数分析範囲を有声音声の倍音構造が十分に保たれている低い周波数領域に限定し，計算量を抑えた。予備的な検討の結果，有声音声の基本周波数の推定には，処理すべき周波数範囲は0 Hz～2000 Hz程度で十分であった。

さらに，本システムは音声認識には利用しないため，音モデル（セントロイド）を推定する必要は無い。極端にいえば，たった1つの音モデルを用いるだけでも構わない。予備的な検討の結果，音モデルの数を1にまで減らしても，基本周波数の推定結果にさほど大きな影響がないことが分かった。そこで，音モデルの数を1つにしてさらなる処理の高速化を図った。

以下の節では，有声区間の検出と，その結果を利用した音声区間の検出方法について，より具体的に述べる。

有声区間の検出　先に述べたように，入力音声の特定の時刻(フレーム)についてPHONOBESTが推定する重み値$\sum_{m=1}^{M} w(F,m)$は，基本周波数候補Fの関数となり，$\sum_{m=1}^{M} w(F,m)$のFに

関する集中度はそのフレームの入力音声の「有声音らしさ」と対応すると考えられる。具体的には，あるフレームでの$\sum_{m=1}^{M} w(F,m)$の値がある特定のFについて大きくその他のFに関しては小さい場合は，そのフレームでの音声信号が倍音構造を持っている可能性が高く，有声音らしいと見なすことができる。逆にあるフレームでの$\sum_{m=1}^{M} w(F,m)$の値がどのFについてもほぼ同じであれば，そのフレームの音声信号は倍音構造をもっていない可能性が高く，有声音らしくないと見なすことができる。

このような考え方に基づき，あるフレームの入力信号の「有声音らしさ」を，$\sum_{m=1}^{M} w(F,m)$のFに関する集中度係数で表わし，これを「有声度」と定義した。有声度が高ければそれだけある特定の基本周波数候補の重み値が高いことを示しており，したがってより有声音らしいことを表している。なお，集中度係数は総和が1になるため，後の閾値処理がしやすいという利点も持っている。

$$\text{有声度} = 1 - \frac{-\sum_{m=1}^{M} w(F,m)\log w(F,m)}{\log(M)} \tag{5}$$

本システムでは，このように定義した入力信号の各フレームでの有声度を閾値処理して，まずそのフレームが有声なのか無声なのかを判定する。そして，有声フレームが数フレーム連続している場合のみ，その区間を有声区間と判定する。

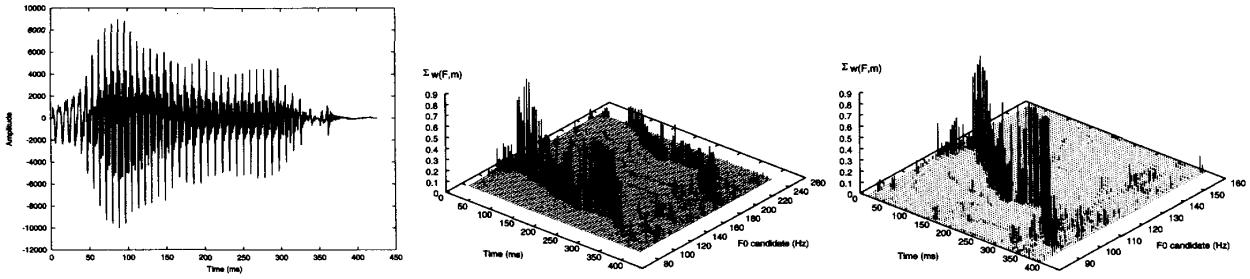
音声区間の検出　有声区間の検出結果から，音声区間を検出する。入力音声は，無声子音から始まる場合と有声子音ないし母音から始まる場合とが考えられる。有声子音または母音から始まる場合は，音声の有声区間の開始点を検出すれば，すなわち音声区間の開始点を検出したことになる。しかし，無声子音から音声が開始される場合は，有声区間の開始点より前に音声区間の開始点が存在することになる。

ここで，本システムでは，「ささやき声」のように入力音声が全てあるいはほとんど無声化している場合は検出の対象から外して考えている。なぜなら，本システムは騒音下での運用を想定して作成しており，騒音下で「ささやき声」のような無声化した音声を検出することはそもそも非常に困難であるし，また，本システムを使うユーザの側も騒音下で「ささやき声」を入力するとは，まず考えられないからである。

さて，有声区間の開始点から音声区間を検出するにあたり，本システムでは入力信号の開始が有声であるか無声であるかに関わらず，有声区間の開始点から若干前の時刻が音声区間の開始点であると仮定して処理を進める。つまり，システムは有声区間の開始点よりも予め設定された適当な時間だけ前に音声区間の開始点を設定する。これは，先に述べたように本システム

は騒音下での運用を想定しており，騒音下で音声の無声部分の開始点を検出することは現実的ではないからである。

音声の終了点を検出する場合も開始点の場合と同様に，検出された有声区間の終了点からある適当な時間だけ後の時刻で音声が終わっていると仮定し，その時刻に音声区間の終了点を設定する。予備的実験の結果，有声区間の開始点から50 ms前に音声区間の開始点を設定し，有声区間の終了点から50 ms後ろに音声区間の終了点を設定することが最適であることが分かった。



(a) テストに用いた音声波形：“ほうえ”　(b) 倍ピッチ処理前　(c) 倍ピッチ処理後

図2　倍ピッチ処理前と倍ピッチ処理後の$\sum_{m=1}^{M} w(F,m)$の値

倍ピッチ成分の処理　予備的実験の結果，PHONOBESTで音声の有声部分を処理した場合，推定された$\sum_{m=1}^{M} w(F,m)$には音声の基本周波数だけでなく，基本周波数の倍の周波数成分(倍ピッチ）も出現するという現象が観察された。具体的にいえば，音声信号の基本周波数がfHzである場合，$\sum_{m=1}^{M} w(F,m)$は，基本周波数（$=f$Hz）に相当するところの値が大きいだけでなく，その倍の周波数（$=2f$Hz）に相当するところの重み値も若干大きいことが分かった。

そこで，$\sum_{m=1}^{M} w(F,m)$のうち倍ピッチに相当する周波数の重み値をその半分の周波数の重み値に加算することにし，$\sum_{m=1}^{M} w(F,m)$の値を修正した。この修正により，$\sum_{m=1}^{M} w(F,m)$において倍ピッチ成分の重みが高くなることによる集中度係数が見かけ上の低下を防止することができた。

図2に，PHONOBESTで処理を行った結果得られた$\sum_{m=1}^{M} w(F,m)$の値を，倍ピッチ処理前の場合と倍ピッチ処理後の場合について示す。音声資料は，男性話者の「ほうえ」という発話音声を用いた。

倍ピッチ処理前の図2(b)を見ると，周波数が180 Hz～240 Hz付近に倍ピッチ成分が出現していることが分かる。倍ピッチ修正処理を行った後には（図2(c)），倍ピッチ成分が基本周波数

成分に吸収され，基本周波数成分の$\sum_{m=1}^{M} w(F,m)$のみが大きな値を取っていることが分かる。

3 性能評価実験

高騒音下での提案手法の性能を評価するため，雑音が重畳されていない音声(クリーン条件)，非定常雑音を重畳した音声（騒音条件：SNR＝0 dB，10 dB)，のそれぞれを用いて，従来法と提案手法の音声区間検出性能の比較評価を行った。

3.1 方法

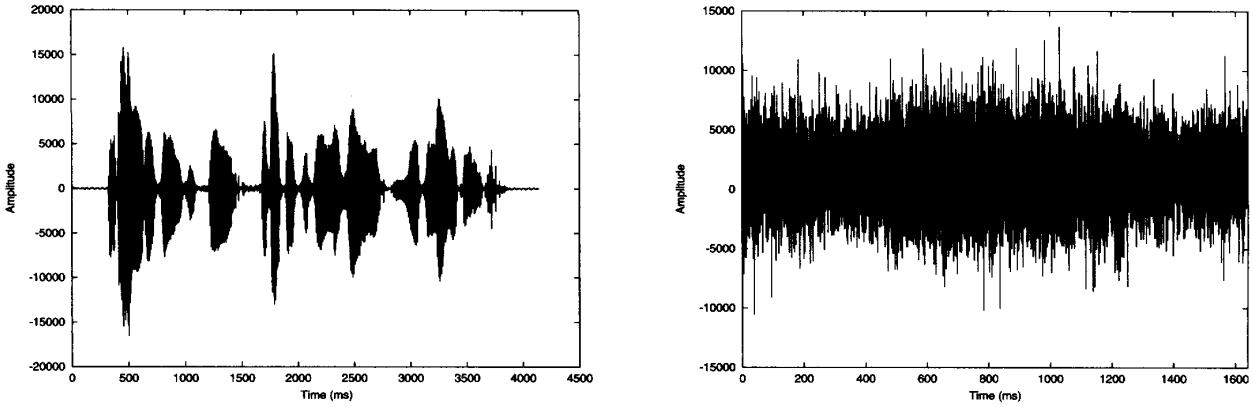
音声資料 ATR提供の多数話者音声データベース・音素バランス文Aセット［1］の50文を男性話者2名（話者ID: M7KENA, M7MAII）が発話した音声データ（合計100文）をテスト音声資料とした。

データベースに収録されている音声資料は，サンプリング周波数16000 Hz，量子化ビット数16 bit，チャンネル数1（モノラル)，big-endianフォーマットで符号化されている。評価実験では，これを，サンプリング周波数は8000 Hz，量子化ビット数は16 bit，チャンネル数1（モノラル)，little-endianフォーマットに変換して用いた。各音声資料は音声区間の切り出し前のもので無音部が含まれており，持続長はおよそ5秒～7秒程度である。文章の一部を表1に，評価実験で音声資料とした音声波形（「あらゆる現実をすべて自分のほうへねじ曲げたのだ」)の例を図3(a)に示す。

表1 評価実験で用いた発話文の一部

発話文番号	発話内容
(01)	あらゆる現実をすべて自分のほうへねじ曲げたのだ。
(02)	一週間ばかりニューヨークを取材した。
(03)	テレビゲームやパソコンでゲームをして遊ぶ。
：	：
(50)	逆境に耐えたこのプロデューサーの作品には， ヒューマニズムが脈々と息づいている。

騒音資料 JEIDA騒音データベース［7］に含まれる工場騒音の一部を騒音資料として用いた。騒音資料の一部の区間の信号波形を，図3(b)に示す。この工場騒音は，ガウス性の暗騒音に加えて，比較的レベルの高い自動車騒音，金属のぶつかり合う打撃音，「ピー」という信号音，



(a) 発話音声波形の例　　(b) 騒音波形の例

図3　実験で用いた発話音声波形と騒音波形の例

などの突発的な騒音が含まれているものであり，非定常性が比較的高い。

実験条件　データベース音声をそのまま用いる条件(クリーン条件)と，JEIDA工場騒音をSNR＝10 dB，0 dBの2つの条件で音声に重畳した条件（騒音条件），という3つの条件を設けた。テスト資料作成のために用いた騒音資料の一部は，十分に長い騒音資料の，ランダムに決められた任意の時点から個々の音声資料と同じ長さの区間を切り出して作成した。音声資料とこの騒音資料の一部を2つのS/N比，SNR＝10 dBとSNR＝0 dBで重畳することにより，テスト資料を作成した。

ベクトル量子化コードブックは，評価実験に用いる音声資料（話者2名×50文の発話音声）を用いてHTK3.0［6］のHQuantツールにより作成した。コードブックサイズ（音モデルの数）は1である。音声データは発話区間のみを切り出してから用いた。なお，音声区間の切り出しには音声データベースに付属する時刻つき音素データを利用した。

3.2　結果および考察

PHONOBESTによる有声区間の検出処理　PHONOBESTに話者2名の50文の発話音声を入力し，音声資料の各時刻フレームごとに，PHONOBESTにより推定された$\sum_{m=1}^{M} w(F,m)$から，式(5)に基づき集中度係数を算出し，そのフレームの有声度とした。PHONOBESTの周波数分析部では，サンプリング周波数8000 Hz，分析窓長32 ms，フレームシフト5 msの設定で入力信号の分析を行った。各時刻フレームでの有声度を，その前後3フレームに関して移動平均し，そのフレームでの有声音らしさの指標とした。フレーム間で移動平均した後の集中度係数に対して閾値処理をし，有声度が閾値を上回ったフレームを有声フレームとした。

有声区間の検出結果を基に，音声区間の検出を行った。本システムでは，有声音声区間の前後に常にある程度の長さの無声音声が存在するものと仮定し，音声区間の開始点を有声区間の開始点よりも若干前の時刻に，音声区間の終了点を有声区間の終了点よりも若干後ろの時刻に設定することで，音声区間を検出した。仮定する無声区間の長さは，50msとした。つまり，PHONOBESTの音声区間検出手順は以下の３つの段階からなる。各数値は予備実験の結果に基づいて最適な結果が得られるような値に設定している。

１．有声フレームの判定

有声度（$\sum_{m=1}^{M} w(F,m)$のFに関する集中度係数）を閾値処理し，閾値以上の値が得られたフレームを有声フレームとする。

２．有声区間の検出

連続して50 ms（10フレーム）だけ有声フレームが連続したらその区間を有声区間とみなす。

３．音声区間の検出

有声区間開始フレームから50 ms（10フレーム）前を音声区間の開始フレームに，有声区間終了フレームから50 ms（10フレーム）後を音声区間の終了フレームとする。

このように検出した音声区間を，音声データベース［１］付属の時刻つき音素データに示された音声区間(以下，「手動検出による音声区間」とする)と比較することにより，PHONOBESTによる音声区間の検出精度の評価を行った。

検出精度の評価は，後藤ら［４］の方法にならい，再現率(recall rate)，適合率(precision rate)，およびF値（F-measure）により行った。それぞれの指標は以下のように定義される。

$$\text{再現率}\ (R) = \frac{\text{正しく検出した音声区間}}{\text{手動検出による音声区間}} \tag{6}$$

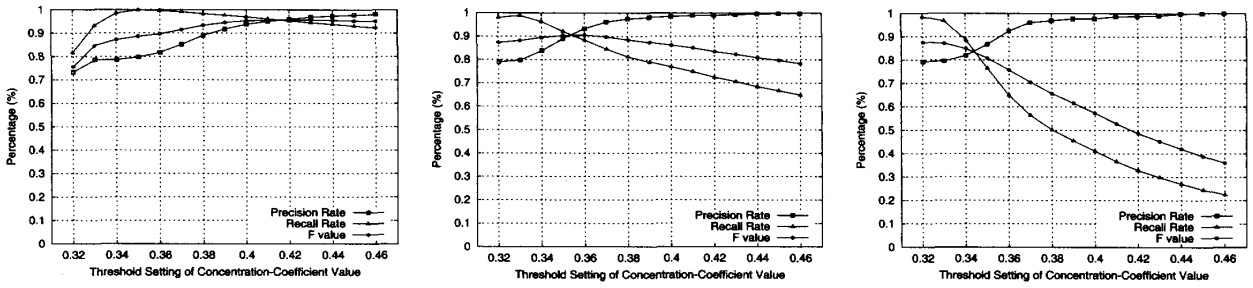
$$\text{適合率}\ (P) = \frac{\text{正しく検出した音声区間}}{\text{検出した音声区間}} \tag{7}$$

$$F\ \text{値} = \frac{(x^2+1)PR}{x^2P+R} \tag{8}$$

ここで，xは再現率と適合率の相対的寄与を決定する係数である。今回は両者を等価に扱い

たいため，$x=1.0$とした。式(8)で定義されるF値が検出精度の高さを表わす指標となる。

また，PHONOBESTによる音声区間の検出精度は$\sum_{m=1}^{M} w(F,m)$の有声度の閾値設定により大きく変化する。そこで，閾値を変化させた場合の検出精度の変化について検討した。図4に結果を示す。



(a) クリーン条件　(b) 騒音重畳条件(SNR＝10 dB)　(c) 騒音重畳条件(SNR＝0 dB)

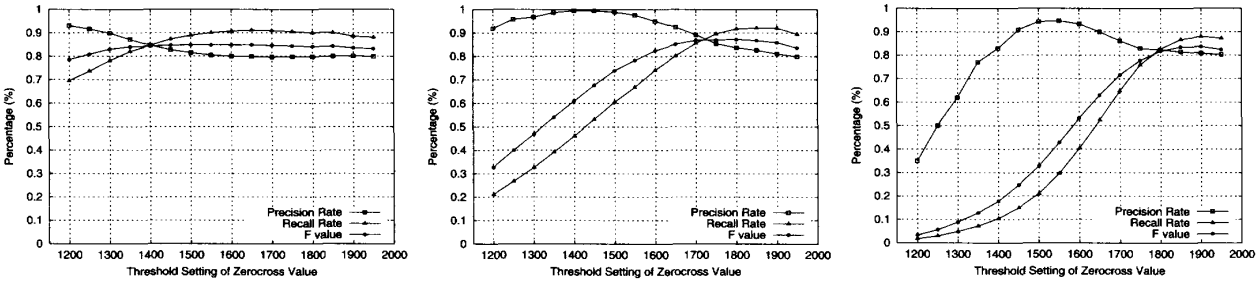
図4　種々の閾値設定でのPHONOBESTの音声区間の検出精度(話者2名×50発話(N=100)の平均値)

図4より，クリーン条件と，SNR＝10 dBの条件では適合率と再現率がほぼ同程度である場合にF値が最大値をとることが分かる。SNR＝0 dBの条件では，閾値設定を上げるにつれ急激に再現率が低下しているために，適合率と再現率がほぼ同程度になるポイントよりも低い閾値設定でF値が最大に近づいている。

表2　最適な閾値設定をした場合のPHONOBESTとゼロクロス法のそれぞれによる音声区間検出精度。数字は話者2名×50発話(N=100)の平均値。[　]内はその標準偏差を表わす。

	テスト条件		
	クリーン	SNR＝10 dB	SNR＝0 dB
		PHONOBEST	
適合率	0.950 [0.029]	0.890 [0.064]	0.820 [0.059]
再現率	0.961 [0.029]	0.920 [0.043]	0.886 [0.059]
F値	0.955 [0.021]	0.903 [0.033]	0.849 [0.039]
		ゼロクロス法	
適合率	0.847 [0.061]	0.854 [0.079]	0.823 [0.073]
再現率	0.847 [0.100]	0.896 [0.079]	0.827 [0.127]
F値	0.843 [0.065]	0.870 [0.053]	0.817 [0.071]

そこで，適合率と再現率が同程度であることを重視して，さらに，F値も大きくなることを基準に最適な閾値設定をするなら，図4より，最適な有声度の閾値の設定はクリーン条件の場合で0.41，SNR＝10 dBの場合で0.35，SNR＝0 dBの場合で0.34程度になると考えられる。これらの最適な閾値設定での音声区間の検出結果を表2に示す。



(a) クリーン条件　(b) 騒音重畳条件(SNR＝10 dB)　(c) 騒音重畳条件(SNR＝0 dB)

図5　種々の閾値設定でのゼロクロス法による音声区間の検出精度（話者2名×50発話（N＝100）の平均値）

従来法による音声区間検出　PHONOBESTによる音声区間の検出が従来法と比較してどの程度優れているのかを評価した。比較する従来法としては，音声波形のゼロクロス値に基づく音声区間検出を選択した。この方法は，ある一定時間内に音声波形の振幅値の正負が逆転する回数（振幅0を横切る回数）が有声部では小さく，無声部または騒音部分では大きくなることを利用して有声区間を検出する方法である。この方法はPHONOBESTの場合と同様に音声信号の有声／無声判定をする方法であるため，このままでは音声区間の検出をすることはできない。そこで，本システムと同様に有声区間の前後に50 msの無声区間が存在するものと仮定して音声区間を検出するように，ゼロクロス値による検出法を修正して性能比較を行った。

ゼロクロス法の場合も，PHONOBESTと同様に音声資料の各時刻フレームでのゼロクロス値の閾値処理を行っている。そこで，ゼロクロス値の閾値を変化させた場合の検出精度の変化について検討した。図5に，ゼロクロス法による推定精度を示す。

図5より，やはり適合率と再現率がほぼ同程度である場合にF値が最大値をとっているものの，適合率と再現率が同程度の値になる閾値設定とF値が最大値をとる閾値設定とは，条件によっては若干ずれている場合がある。

PHONOBESTの場合と同様に，適合率と再現率が同程度であることを重視して，さらにF値も大きくなることを基準に最適な閾値設定をするとすれば，図5より，最適な閾値はクリーン条件の場合で1400，SNR＝10 dBの場合で1750，SNR＝0 dBの場合で1800程度になると考え

られる。これら最適な閾値設定での音声区間の検出結果を表2に示す。

PHONOBESTによる音声区間検出結果と従来法による音声区間検出結果の比較　表2から，適合率と再現率の両者を考慮した指標であるF値によりPHONOBESTと従来法の音声区間検出の精度を評価した場合，どの実験条件でもPHONOBESTのF値がゼロクロス法のF値を上回る精度を出していることが分かる。PHONOBESTのゼロクロス法に対する優位性は，クリーン条件では約11%と大きく，SNR＝10 dB条件とSNR＝0 dB条件では，それぞれ約3%と小さい。しかしながら，分散分析による検定を行ったところ，PHONOBESTとゼロクロス法の間のF値の差はクリーン，SNR＝10 dB，SNR＝0 dBのすべての条件で有意であった（それぞれ，$F(1,49)=477.96$, $p<.01$；$F(1,49)=36.79$, $p<.01$；$F(1,49)=16.90$, $p<.01$）。

また表2より，F値の標準偏差（SD）の値をPHONOBESTとゼロクロス法の場合とで比較すると，PHONOBESTのF値のSDは，ゼロクロス法のF値のSDの1/2から1/3程度と非常に小さくなっている。このことは，PHONOBESTがゼロクロス法と比較して安定して音声区間の検出を行っていることを意味している。

以上より，クリーン音声に対しても，また騒音が重畳された音声に対しても，PHONOBESTは従来のゼロクロス法と比較してより高い精度で（F値の平均値），またより安定して（F値のSD）音声区間の検出を行うことができることが確認された。

PHONOBESTによる音声区間検出がうまくいかない場合　無音区間が比較的短い音声資料（話者M7KENA. SD. A，文01）と無音区間が長い音声資料（話者M7KENA. SD. A，文12）を入力した場合の，本システムによる音声区間推定の例を図6，図7にそれぞれ示す。また，それらの検出精度を，表3，表4にそれぞれ示す。

これらの結果から，まずクリーン音声の場合は無音区間が短い音声資料（図6，表3）の場合も，無音区間が長い音声資料（図7，表4）の場合も非常に正確に音声区間が検出できていることが分かる。しかし騒音が重畳された場合は，重畳される騒音のレベルが大きくなるほどどちらの音声資料の場合も音声区間の検出は不正確になっている。そして騒音が重畳されることによる音声検出精度の悪化は，無音区間が比較的長い音声資料（図7，表4）の場合に，無音区間が比較的短い音声資料の場合（図6，表3）よりも顕著であることが分かる。

以上の結果を総合的に考えると，PHONOBESTの音声区間検出精度が悪化した原因は，音声信号が存在せず騒音のみが存在している区間での誤検出が大きな要因となっていると考えられる（図7）。評価実験では，PHONOBESTの有声度の閾値処理において閾値を固定して音声

区間を検出しているが，入力信号に応じて閾値を適応的に変化させることで音声区間検出性能のさらなる向上を図ることが可能かもしれない。これは今後の課題である。

表3　無音区間が比較的短い音声資料（話者 M7KENA. SD. A，文01）の検出精度

	テスト条件		
	クリーン	SNR＝10 dB	SNR＝0 dB
適合率	0.986	0.986	0.942
再現率	0.931	0.878	0.776
F値	0.958	0.929	0.851

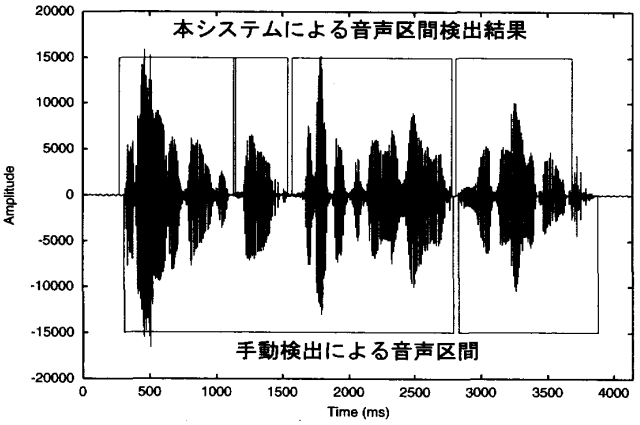
表4　無音区間が比較的長い音声資料（話者 M7KENA. SD. A，文12）の検出精度

	テスト条件		
	クリーン	SNR＝10 dB	SNR＝0 dB
適合率	0.944	0.672	0.635
再現率	0.968	0.971	0.921
F値	0.956	0.794	0.752

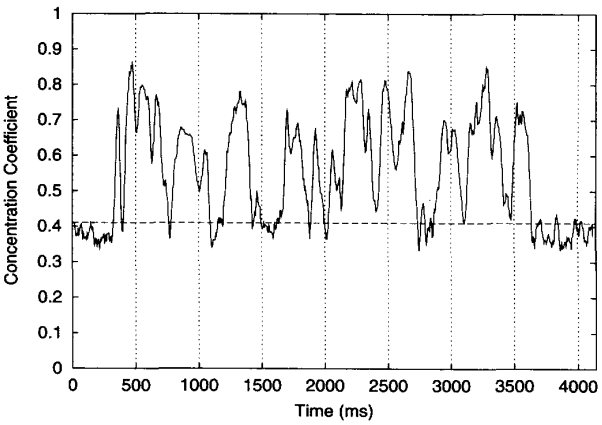
4　まとめと今後の展望

㈶九州システム情報技術研究所にて開発された騒音下音声認識システム PHONOBEST を拡張し，入力信号に音声信号が含まれる区間の検出を行う処理を作成し，その性能に関する定量的評価を行った。複数の男性話者による発話音声を用いて，提案手法と従来法の音声区間検出性能を比較した。

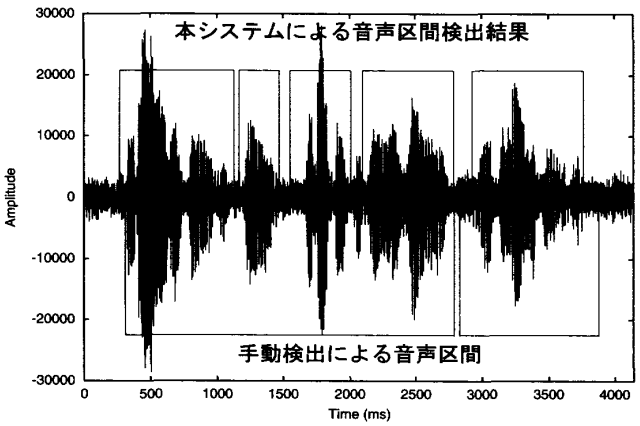
この結果，騒音が重畳されていないクリーン音声に対しては従来法と比較して約11％の検出性能の向上が示され，また，音声に対して比較的非定常な性質を持つ工場騒音を重畳した場合でも，従来法と比較して SNR（signal to noise ratio）が10 dBの条件と0 dBの条件でそれぞれ約３％の音声区間検出性能の向上が示された。さらに本システムは，クリーン音声に対しても騒音が重畳された音声に対しても，従来法と比較してより安定して音声区間の検出を行うことができることが示された。



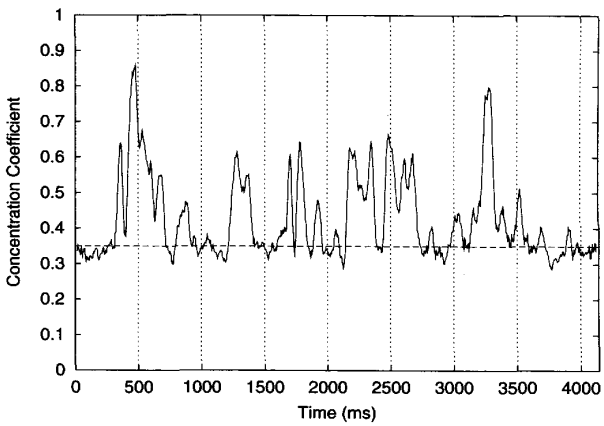
(a) クリーン条件・検出結果



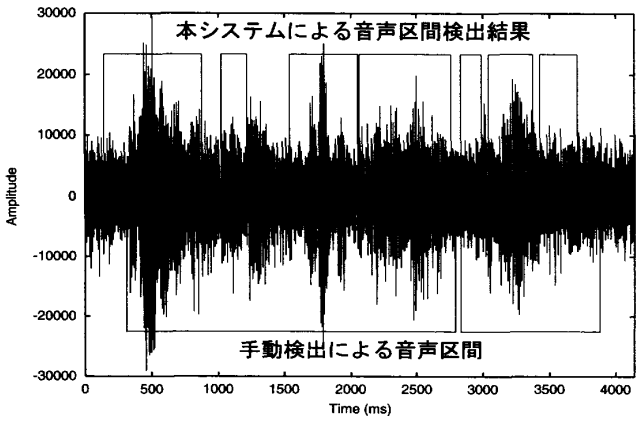
(b) クリーン条件・有声度



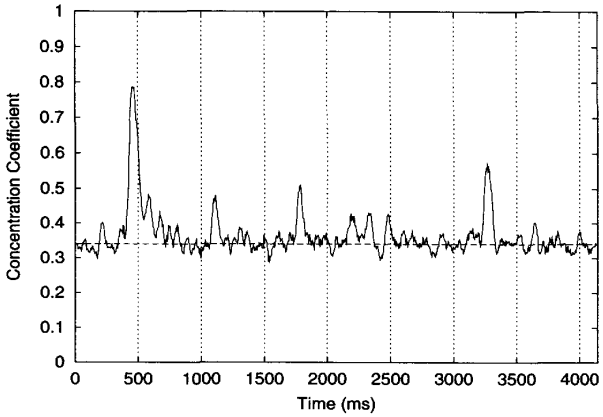
(c) SNR＝10 dB条件・検出結果



(d) SNR＝10 dB条件・有声度

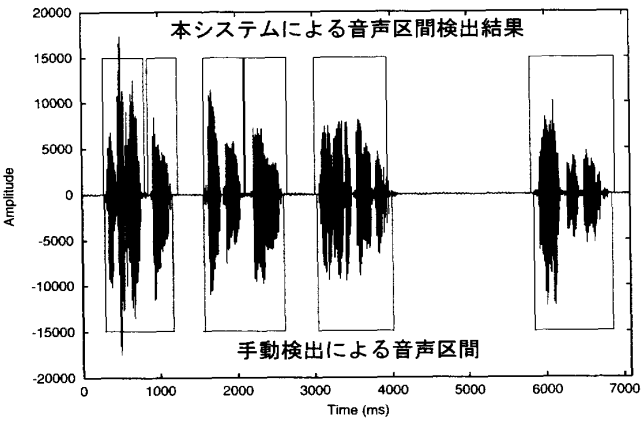


(e) SNR＝0 dB条件・検出結果

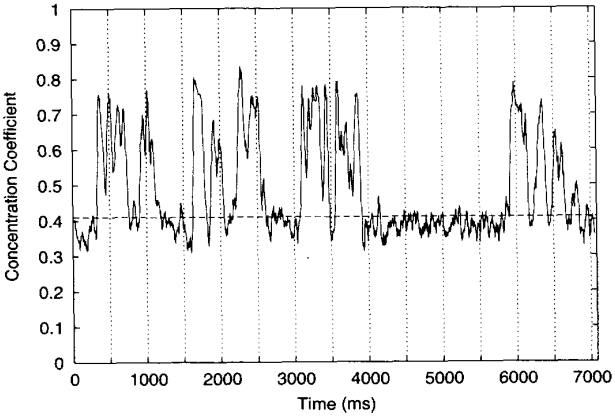


(f) SNR＝0 dB条件・有声度

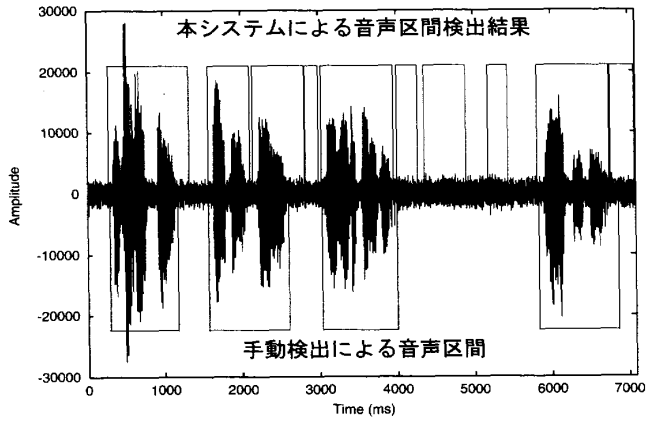
図6　無音区間が短い音声資料（話者 M7KENA. SD. A，文01）の音声区間検出結果と有声度の推移



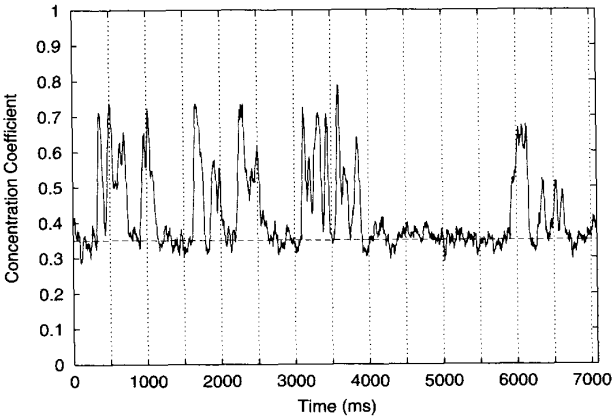
(a) クリーン条件・検出結果



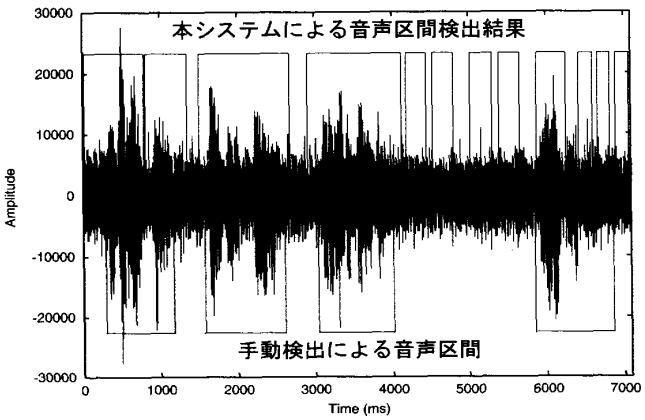
(b) クリーン条件・有声度



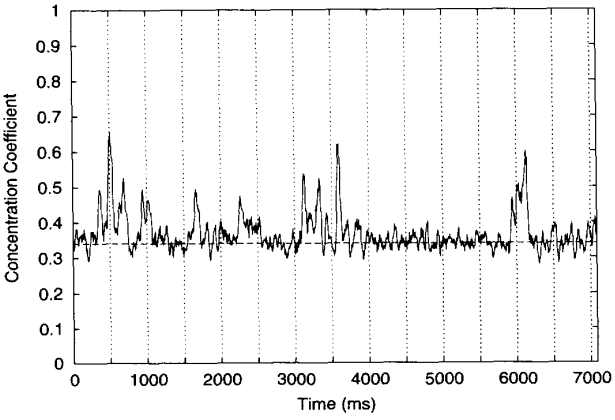
(c) SNR＝10 dB条件・検出結果



(d) SNR＝10 dB条件・有声度



(e) SNR＝0 dB条件・検出結果



(f) SNR＝0 dB条件・有声度

図7　無音区間が長い音声資料（話者 M7KENA. SD. A，文12）の音声区間検出結果と有声度の推移

しかしながら，今回開発したシステムは処理速度が非常に遅く，少なくともPCレベルの性能のコンピュータではリアルタイム処理できないことが欠点として挙げられる。それに対して，従来法（ゼロクロス法）ではほぼリアルタイムの処理が可能である。このシステムを実環境で運用するためには，今後，処理速度を大幅に向上させる必要があろう。

また，PHONOBESTの有声度の閾値を入力信号に応じて適応的に変化させることは，音声区間検出性能を向上させるために有効な方法かもしれない。今後は，これらの点を改善することが必要となるであろう。

謝辞

本研究は，㈶九州システム情報技術研究所からの受託研究補助金（2002年4月22日～2002年6月5日）により行われた。㈶九州システム情報技術研究所の勝瀬郁代研究員（現・近畿大学産業理工学部）には，本研究の遂行にあたり様々な有益なアドバイスを頂いた。

参考文献

[1] ATR㈱「多数話者音声データベース・音素バランス文」（CD-ROM，APP-BLA），ATR㈱国際電気通信基礎技術研究所，京都，1997年.

[2] Boll, S. F., "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. ASSP-27, 1979, Pp.113-120.

[3] 後藤真孝「リアルタイム音楽情景記述システム：全体構想と音高推定手法の拡張」『情報処理学会音楽情報科学研究会研究報告』2000-MUS-37-2，第2000巻第94号，2000年，9～16ページ。

[4] 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」『電子情報通信学会論文誌』D-Ⅱ，第J83-D-Ⅱ巻第11号，2000年，2330～2340ページ。

[5] Goto, M., "A predominant-F 0 estimation for CD recordings: MAP estimation using EM algoritm for adaptive tone models", Proc. of ICASSP 2001, 2001, Pp. V-3365-3368.

[6] Hain, T., Woodland, P. C., Evermann, G., & Povery, D., "The CU-HTK March 2000 HUB5E transcription system", Proc. of Speech Transcription Workshop 2000, 2000.

[7] 板倉秀一「騒音データベースと日本語共通音声データDAT版」『日本音響学会誌』第47巻第2号，1991年，951～953ページ。

[8] 勝瀬郁代，菅野禎盛「PreFEstの騒音下音声認識への応用」『日本音響学会講演論文集』1-Q-14，2001年10月，167～168ページ。

[9] 勝瀬郁代，菅野禎盛「PHONOBEST:‘期待’に基づく音韻推定処理を組み込んだ雑音に頑健な音声認識システム」『信学技法』SP2002-127，2002年，17～24ページ。

[10] Masuda-Katsuse, I., "A new method for speech recognition in the presence of non-stationary, unpredictable and high-level noise", Proc. of 7th European Conference on Speech Communication and Technology (Eurospeech 2001), 2001, Pp.1119-1122.

[11] Masuda-Katsuse, I., & Sugano, Y., "Speech estimation biased by phonemic expectation in the presence of non-stationary and unpredictable noise", Proc. of CRAC workshop, 2001.

[12] 鹿野清宏，中村哲，伊勢史郎『音声・音情報のディジタル信号処理』昭晃堂，1987年。

[13] 渡部生聖，山田武志，北脇信彦，浅野太「環境音モデルとHMM合成を用いた音声区間検出の検討」『電子情報通信学会技術研究報告』第100巻第520号（NSC2000 27-46），2000年，55～60ページ。