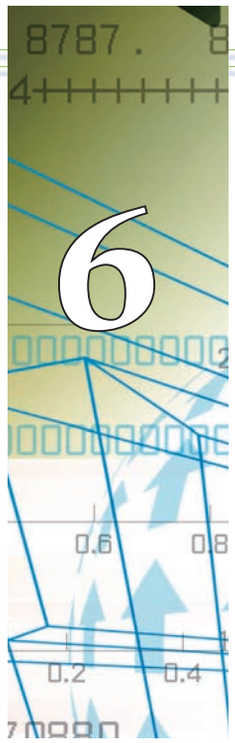

特集 音声情報処理技術の最先端

自動車の中での音声認識

武田 一哉
名古屋大学情報科学研究科
takeda@is.nagoya-u.ac.jp

自動車内での利用を目的とした音声認識技術の現状と研究動向を解説する．まず，大規模な実データを用いた実験結果を用いて，車内音声認識の性能の現状を示し，自動車内での音声認識の主たる問題点（雑音の混入と走行環境の変動）を指摘する．次に，車内での音声認識性能を向上させるための，さまざまな音響信号処理手法（単一／複数チャネルの雑音抑圧手法，雑音への適応手法，発話区間検出法，画像情報を併用する手法など）を概説する．さらに，車内での音声認識の評価方法（AURORA3標準評価法，利便性と安全性の評価研究例）について議論する．最後に，車内音声認識の研究資源と新しい研究動向を紹介する．

■自動車内での情報インタフェース

カーナビゲーションシステムの目的地設定や制御を中心に，自動車内での音声認識の実用化が進んでいる．月間30万台程度出荷されているカーナビゲーションシステム[1)]の40％程度が音声認識機能を搭載していると考えられている．現状では，音声入力は補助的な操作機能を果たすことが普通であり，音声入力が利用される頻度はあまり高くないと想像されている．多くのカーナビゲーションシステムは，あらかじめ定められた語彙内の単語が孤立して発声されることを想定しており，利用者にとって必ずしも自然な入力インタフェースではないこと，そして自動車内の音声認識性能がいまだ十分高くないこと，などがその原因と考えられる．

しかし，無線，インターネット，地図情報などの技術が急速に発展する中，移動中の自動車内で，情報サービスを利用する要望は大きく，自動車運転中でも利用できるインタフェースとして，音声認識に対する期待は増している．これらの背景の中，自動車内での音声認識技術を高度化するべく，さまざまな研究開発が進められている．

■自動車内における音声認識の問題点

自動車内で運転以外の作業を行うためには，いわゆる「ハンズフリー」な状態でシステムを動作させることが必要になる．音声認識の場合，マイクロホンを口元に近づけることが困難である一方，接話マイクを備えたヘッドセットを装着することは運転者の負担が大きいことから，離れた場所（運転者の口元から，40～50cm程度）に設置された遠隔マイクロホンで受音を行うことになる．このため走行中の自動車内では，雑音や残響音に対する音声の比率（SN比）が低下した劣化音声を認識する必要がある．通常の市街地走行中に，遠隔マイクで受音された音声は，接話マイクで受音する場合に比べて平均で10dB程度SN比が劣化している．図-1は，走行中の自動車内で運転者が孤立発声した50単語の単語音声を認識した結果を示している．接話マイクロホンで受音された音声は，周囲の状況にかかわらず90％以上の高い精度で認識可能であるのに対し，遠隔マイクロホンで受音された音声の認識精度は雑音の影響を受け

て劣化している．

同様に図-2には，自然な音声インタフェースを提供するために不可欠な，大語彙連続音声認識の自動車内での性能例を示す．実際に市街地走行中の自動車内で，運転者がレストラン案内サービスにアクセスするために発声した文音声を，接話マイク，遠隔マイクそれぞれで受音し，話者ごとの平均SN比と音声認識性能（単語正解精度）を図示している．音響モデルと言語モデルには，400名の車内音声で学習した隠れマルコフモデル（HMM）とトライグラム言語モデルをそれぞれ用いた．認識の対象とする語彙の規模は1,800語から5,000語である．図から，10dBのSN比の劣化が5％の認識精度の低下におおよそ対応していることが分かる．しかし，同程度のSN比の音声であっても，得られる認識性能が話者によって大きくばらついており，平均的な単語正解精度も遠隔マイクを用いた場合には70％を下回っている．

このため，自動車内の音声認識性能を向上させるためには，まず雑音の影響を取り除くとともに，多様な環境の変動に頑健な音声認識技術を確立することが重要である．

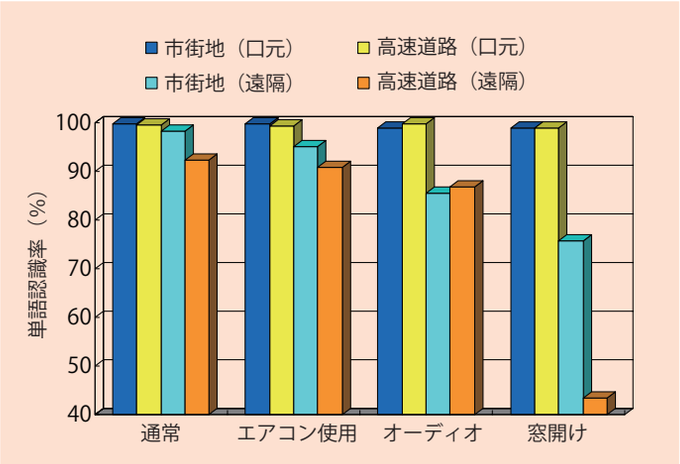

図-1　小語彙の単語発声に対する車内音声認識性能

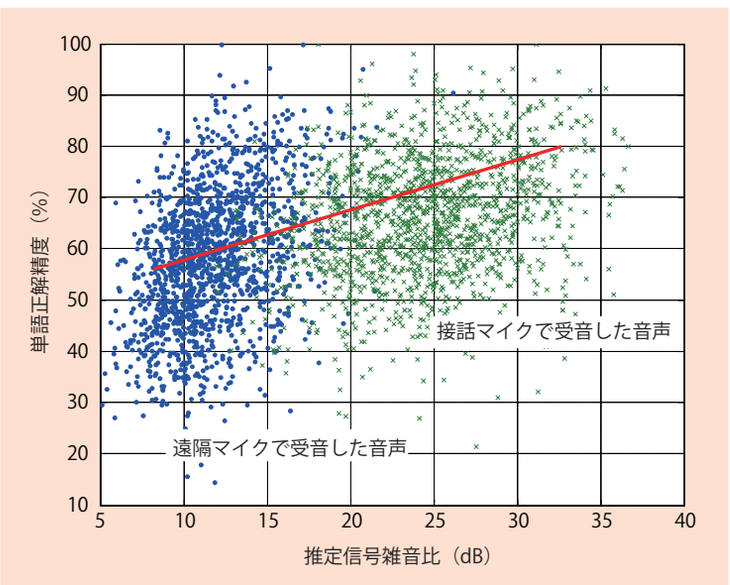

図-2　SN比のばらつきと大語彙連続音声認識に対する車内音声認識性能

■車内音声認識のための音響信号処理

走行自動車の車内騒音

自動車内の音声処理の最も大きな問題の1つに車内雑音の存在がある．走行中の自動車の車内には，エンジン音のほか，いわゆる走行音，風きり音，ラジオ・カーステレオの音，エアコンのファンの音，同乗者の声，などさまざまな雑音が存在する．走行音も定常ではなく，加速，追い越し，対向車とのすれちがいなど，走行状態に応じて変化する．図-3は，4種類の自動車内で走行中に収録された1時間の車内音を人間が聞き取り，聞き取れた雑音イベントの回数を数えた結果を示している．道路の継ぎ目の通過に伴うバンプや，追い越し，対向車とのすれ違いなど，定常走行時においても，雑音イベントとして聞き取ることのできるさまざまな非定常音が観測されることが分かる．

特別な雑音イベントが発生していない定常走行時に，運転席の天井に取り付けられたマイクロホン（無指向性，話者の口元までの距離40cm程度）で収録された車内雑音のスペクトルを図-4に示す．定常走行（アイドリング）時の車内雑音のパワーは，低周波数成分に集中している．一方，エアコン動作時，窓開け時，オーディオ再生時，の車内騒音は2kHz以上の高周波成分を持ち，音声認識の大きな障害になる．

車内音声認識のための雑音除去手法

車内の雑音を除去するためにさまざまな音響処理手法が用いられる．以下に代表的な雑音除去手法について概説する．

(1)低周波成分の除去

図-4に示したとおり，定常走行時の車内雑音のパワー

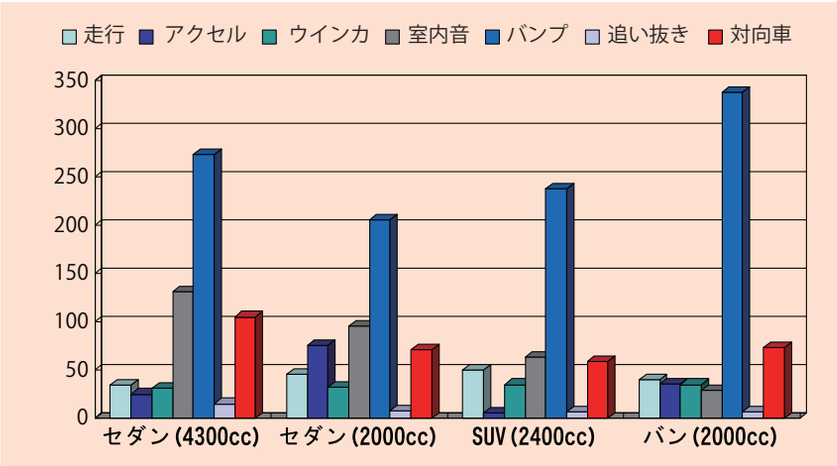

図 -3　定常走行中の自動車内の雑音イベント（1 時間に聞き取れた回数）

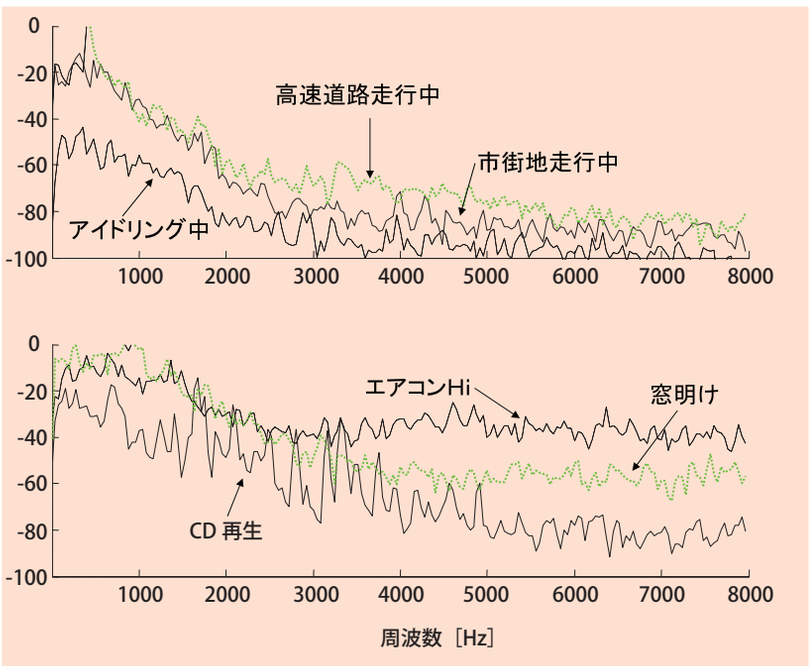

図 -4　車内雑音の平均スペクトル

は低周波成分に集中していることから，自動車内の音声の認識は，100～200Hz以下の低周波成分を除去した後に行うことが一般的である．

(2)（準）定常雑音成分の除去（スペクトル減算法）

一般に音声信号と雑音信号は独立であり，雑音が重畳した音声のパワースペクトルは，音声信号と雑音信号それぞれのパワースペクトルの和で近似的に計算できる．雑音のパワースペクトルは音声が発声されていない区間の信号から比較的容易に推定でき，音声信号に比べて時間的変化が緩やかであるため，雑音のパワースペクトルを逐次的に推定することは，音声のパワースペクトルを推定することに比べて容易である．このことを利用して，雑音が重畳した音声のパワースペクトルから雑音のパワースペクトルの推定値を減算する，スペクトル減算に基づく雑音抑圧法が知られている．他の多くの音声認識応用システムと同様に，スペクトル減算処理は自動車内の雑音抑圧にも有効である．

(3)複数マイクロホンを利用する方法

音声認識の対象を運転者の発声に限定すれば，車内では音声の音源位置がほぼ固定されていると考えることができ，運転者の方向に指向性を持たせた受音を行うことで，運転者以外の方向から到来する雑音の影響を低減

することができる．そこで，マイクロホンアレーを用いた指向性受音を，自動車内の音声認識に適用する方法が研究されている．特に，遅延和アレーと減算アレーを組み合わせた適応ビームフォーマ（図-5）は，オーディオ再生時や窓開け時のように，主たる雑音源が空間的に局所化される場合に効果が大きい．しかし，複数マイクを用いる手法，特に適応的な手法には，システムの複雑さや計算コストが大きいという難点があり，広く実用に供されるには至っていない．

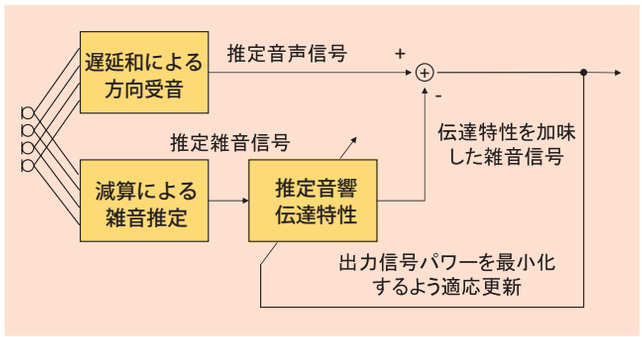

図-5　適応的マイクロホンアレー
（Griffith-Jim 型ビームフォーマ）の構成

(4)ブラインド信号分離の応用

認識対象音声と車内雑音の音源が，受音位置に対して同一方向に存在する場合がある．運転席前方にマイクロホンを設置した場合，運転席側の後部座席からの音声がその典型的な例となる．この場合，受音の指向性を雑音除去に利用することができない．またアプリケーションによっては，認識対象を運転者の音声に限定することができない場合もある．そこで，独立成分分析や，多マイク受音信号の周波数成分ごとの入射角度を推定する方法（SAFIA）など，混合信号のブラインド分離を車内音声の認識に利用する研究も行われている．

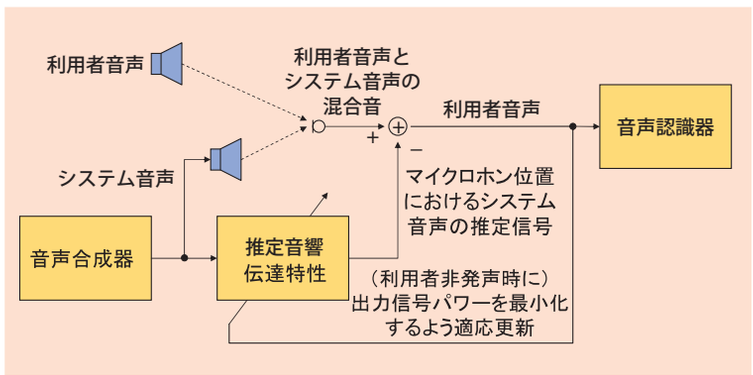

図-6　バージインを実現するための案内音声の消去

(5)雑音への適応

多くの車内音声認識システムは，隠れマルコフモデル（HMM）に基づく音響モデルを用いている．走行に伴い音響環境が変化する状況では，単一の音響モデルを用いるのではなく，複数のモデルを利用し分けることも有効と考えられる．このような環境へのモデル適応技術のうち自動車内で特に重要な技術として，走行雑音に応じて複数の無音モデルを選択的に利用する方法，走行速度別の音響モデルを作成し選択的に利用する方法，などが研究されている．これら音響モデルを適応的に用いる方法により，高雑音下で発声された音声が，静粛時とは異なったスペクトル特徴を呈する「ロンバート効果」の影響を補正することも期待できる．

発話検出と音響伝達特性の推定

マイクロホンと離れた位置から音声を入力する遠隔音声入力環境では，音声入力の検出も容易ではない．音声認識システム自身が生成した（合成音声などによる）案内音声と，利用者の発声とを音響信号だけから区別することが困難なためである（システムが案内音声を発声している最中に，利用者が「割り込んで」発声することはバージインと呼ばれ，バージインを実現することは，対話型の音声インタフェースに不可欠である）．利用者の音声を最も確実に検出する方法は，利用者が発話を開始する前に「発話開始ボタン」を押下し音声認識システムに発話を通知する方法である．しかし，発話開始後にボタンを押す，ボタンをまったく押さずに発話する，などボタン操作自体の確度も完全ではない．

音響信号を用いた発話検出の精度を向上させるために，音響エコーキャンセラを応用して，システムの案内音声を音響的に消去する方法が知られている（図-6）．この方法は，音声出力用スピーカから，受音マイクロホンまでの音響伝達特性を逐次推定し，遠隔マイクで受音された音声からシステムの案内音声のみを除去することで，発話の検出精度を向上させることができる．

映像情報の併用

さまざまな雑音が混入する音響信号に対して，走行車内で撮像した画像信号では，走行に伴う劣化は少ない．車両の振動に伴う画像の揺れ，走行条件によって変化するコントラストの変化などが主たる劣化である．そこで，画像情報を補助情報として用いることで，自動車内の音声認識精度を向上させる方法が研究されている．画像・音声双方の特徴量の時間系列を，従来の音声認識とほぼ

同様にHMMを用いてモデル化し，音声認識を行う手法が研究の中心である．

映像からより高次の情報を抽出し音声認識に利用する方法も検討されている．運転者の顔映像から口の位置や動きを推定し，推定された口の位置に基づいてマイクロホンアレーの指向特性を制御する方法や，口の動きから発話の有無を判定して発話検出の精度を向上させる方法，などがその一例である．

■車内音声認識の評価

音声認識性能の評価標準

携帯電話を利用してネットワーク上で分散して受音・音響処理された音声を，ネットワーク上のサーバで認識処理する「分散音声認識アーキテクチャ」が提案されている．この分散音声認識アーキテクチャのための受音・音響処理技術を評価する枠組みとして，AURORA[2]と呼ばれる国際的な活動が行われている．AURORAでは，さまざまな環境変動要因を組み込んだ評価用データと認識の基本モデル・アルゴリズムを提供しているが，そのうちAURORA3と呼ばれるサブセットは，車内で収録された音声を評価音声としている．AURORA3では，マイク位置（近接マイクと遠隔マイク）,走行条件，の2つの要因で音声認識の難易度（学習時と認識時の環境の一致度）を定め，難易度ごとにさまざまな音響処理アルゴリズムの比較をすることができる．

このような標準的な評価基盤は，与えられた条件の下でさまざまな処理手法を網羅的に比較し，ピーク性能を向上させるためにきわめて効果的であるが，標準データには含まれない変動要因に対する頑健性に対して，十分注意を払うことが重要である．

利便性と安全性

音声認識の主たる目的が運転と並行した機器操作・情報アクセスである以上，自動車内での音声認識システムの性能には，少なくとも利便性と安全性を含む，多面的な評価が必要となる．脇田らは，運転中にLEDの点灯位置を判定回答する，「LED刺激反応タスク」と音声認識システムの利用を運転に並行して行い，音声認識の課題達成時間（利便性）と，LED刺激への反応時間（安全性）の計測を行った[3]．音声認識システムが受理する言い回しには，定型文入力（「3時ごろの市役所から名古屋港まで」），単語列挙（「3時ごろ　市役所　名古屋港」），およびスロットフィリング（時刻を言って下さい→「3時頃」．出発地点を言って下さい→「市役所」等）が用いられた．LED刺激への反応時間は，反応時間が1秒を超える反応遅れが生じた割合で評価をしている．

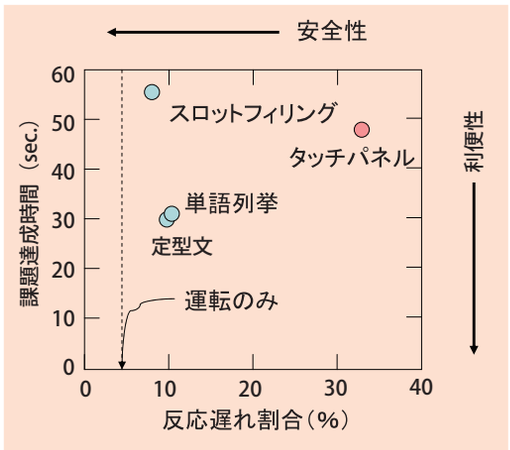

図-7　安全性と利便性からの車内音声認識システム評価

実験の結果を図-7に示す．音声入力を用いたインタフェースは，タッチパネルを用いた視認と手操作に基づくインタフェースに比べて，反応遅れの割合が3分の1に減少していることが分かる．この割合は，音声認識を併用しない場合の反応遅れの割合のほぼ倍である．音声認識システムに対する3種類の言い回しの間で，反応遅れの割合に大きな差異は認められなかった．一方課題達成時間は，単語列挙型，定型文入力の間に大きな差異はなかった．冗長なシステム発話を伴うスロットフィリング型では，課題達成に必要な時間が約50秒と他の入力型の2倍程度であり，タッチパネルよりも長い時間を課題達成に必要としている．

これらの実験結果は合理的であり，反応遅れ割合，課題達成時間という2つの指標が，自動車内における音声認識システムの評価に有効であることが示唆される．しかし，運転者，運転状況，アプリケーションシステムなどの多様な環境下で安定した性能評価を行うためには，さらに研究が必要である．

■研究資源の充実と今後への期待

統計的な手法を基本とする近年の音声認識の研究開発には，実世界の分布に忠実でありかつ巨大なバリエーションを内包する音声データベースが不可欠である．しかし，残念ながら，一般に入手可能な大規模車内音声データベースは多くない．実環境下における車内音声データの収集には，多大なコストが必要なためである．大規模な実データの収集の先駆けとして，米国Colorado大学では500名規模の車内音声データの収集が行われたが，

図 -8　車内音声・行動データベースのブラウザ画面

そのデータは一般に公開されていない．

筆者ら名古屋大学のグループでは，車内12点に分散配置されたマイクロホン（4素子からなるアレーマイクロホンを含む）により車内音声を収録するとともに，3カ所のカメラで撮像された映像および運転信号（アクセル・ブレーキ踏力，速度，エンジン回転速度，ハンドル角度）を同期して集録（図-8）する大規模な車内音声・運転行動データの収集を行い，延べ800人の運転者について実走行中の音声を収集した[4]（本データベースの一部は中部TLOから販売されている[5]）．収録されている音声は，音素出現のバランスがとれた文の読み上げ，レストラン検索に関する対話など1人当たり30分程度であり，すべての発話に書き起こしテキストが付与されている．本データベースを利用して作成されたHMM音響モデルは，情報処理学会傘下の連続音声認識コンソーシアム（CSRC）より会員向けに配布され研究基盤として利用されている．さらに，分散マイクロホン出力を統合する雑音抑圧方法，統計的手法に基づく対話文生成方法，映像と音声を統合した発話区間検出方法，運転行動に基づく個人性認証，など新しい車内音声認識およびその関連技術の研究が当該データベースを用いて進められている[6]．

自動車内での音声認識は，我が国が高い国際競争力を持つ工業製品である「自動車」に，高度な情報・通信技術を集積するために重要な戦略的技術である．自動車内での音声認識技術を発展させ，多様な移動環境下で高度な情報インタフェースを提供することは，「時間・空間に制約されないライフライン」（総合科学技術会議意見具申：『情報通信研究開発の推進』より）の構築においても重要な研究課題であり，今後の進展が期待される．

参考文献

1）（社）電子情報技術産業協会（JEITA）ホームページより，http://www.jeita.or.jp/japanese/stat/shipment/2004/ship_04.htm
2）Hirsch, H. G. and Pearce, D.: The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions, ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium", Paris, France, September.
3）脇田敏裕，寺嶌立太，小島真一，清水　司，本郷武朗：運転中情報機器操作性の評価法，情報処理学会論文誌，42, 1762 (2001).
4）Kawaguchi, N., Takeda, K. and Itakura, F.: Multimedia Corpus of in-car Speech Communication, Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, Vol.36, pp.153-159 (2004).
5）（財）名古屋産業科学研究所，中部 TLO，http://www.ctlo.org/
6）Abut, H., Hansen, J. and Takeda, K. (eds.) : DSP in Vehicular and Mobile Environments, Kluwer (2004).

（平成16年7月13日受付）