Vol.2010-SLP-82 No.4
2010/7/23

車載音声の解析と評価 ～アレイマイクロフォンとスペクトルサブトラクションの融合～

藤澤大希[†]　堀江諒[†]　上野聡[††]　畑岡信夫[†]

カーナビでの音声認識を用いた音声インタフェースの利用向上を目指して，音声認識の頑強性の観点から車載音声の基礎的な解析と認識評価実験を行った．具体的には，アレイマイクロフォン(Array Microphone)を利用して収集した車載音声のSN比の評価やエンジン音の除去を目的としたlow powerカット処理の効果を確認し，最終的にはアレイマイクロフォン加重とスペクトルサブトラクション(SS: Spectral Subtraction)を融合した方式により認識率の向上を実現した．

Analysis and Evaluation of In-Car Speech - Combination of Array Microphones and Spectral Subtraction -

Daiki Fujisawa[†], Ryo Horie[†], Satoshi Ueno[††], and Nobuo Hataoka[†]

In this paper, aiming to a real use of speech recognition for car navigation systems, we analyze and evaluate in-car speech collected in running cars using 7 (seven) array microphones and one head-set microphone. First, we analyze in-car speech from the viewpoint of SN ratio of each microphone and original ASR (Automatic Speech Recognition) accuracy. Second, we evaluate in-car speech from the viewpoints of ASR accuracies after processing a noise reduction method integrating the summation of array microphones and Spectral Subtraction (SS) to obtain ASR accuracy improvement.
Keyword: Speech Interface, ASR (Automatic Speech Recognition), Array Microphones, SS (Spectral Subtraction), Car Navigation System, Car Telematics

1. はじめに

現在，車でのカーナビゲーション搭載率は約80%であり，最近では，カーナビすべてに音声認識付き音声インタフェースが搭載されている．しかし，音声認識機能を利用している人は少ない．その理由として，音声認識技術に問題があるといえる．音声認識技術は，入力された音声を文字に変換し，規定された単語のどれかに判定する技術である．現状の問題として，音声認識技術の問題とインタフェースとしての問題の二つがあり，具体的には，

・実環境で動作しない(車載雑音，発話者以外の声で誤動作等)
・使用方法が分かりづらい(何を言っていいか分からない，話し始めのタイミング)
・語彙外発話の際，正しく認識されない

等がある．その結果，使い方が難しく性能の悪いスイッチでしかないことが想定される．これらの問題を解決するために，本研究では認識率という点に着目し，車載音声の雑音除去を目的にして研究に取り組んでいる．研究の経緯は，早稲田大学IT研究機構音声技術実用化研究所再委託で，(株) 日立製作所が収集した車載音声[1]をデータとし，アレイマイクロフォンの各SN比と音声認識率の関係を調査し[2]，スペクトルサブトラクション(SS: Spectral Subtraction)の効果とアレイマイクロフォン(Array Microphone)の加重効果[3],及びアレイマイクロフォン加重とSSとの融合の効果に関して報告する[4]．

2. 車載応用での音声インタフェースの課題

2.1 ネットワーク応用でのシステムイメージ

IT（Information Technology）がユビキタス・モバイル環境に浸透している現在，ネットワークを介した情報収集において，知的なインタフェースが重要な要素となる．モバイル環境での情報アクセスには，音声を使用したインタフェースがキー技術である．特に，携帯端末やカーナビなどのようにキーボードが使用できない環境では，音声利用（音声認識と音声合成）は必須である．携帯端末（PDAs: Personal Digital Assistants）や携帯電話，さらにはモバイルPCs（Personal Computers）がインターネットなどのネットワークを介して，WEBサーバに接続され，使用者は「いつでも，どこでも，誰でも」が簡単に必要な情報を入手できる環境が整備されてきている．

カーテレマティクス（Car Telematics）では，カーナビ端末や携帯電話を介して，新しい情報サービスを享受できる環境になっている．最近では，InternetRSS（Rich Site

[†] 東北工業大学　工学部　知能エレクトロニクス学科
Tohoku Institute of Technology, Department of Electronics and Intelligent Systems
[††] 株式会社ナカヨ通信機
Nakayo Telecommunications Incorporation

ⓒ 2010 Information Processing Society of Japan

Summary）フィードなどの新しいサービスが台頭している．サービスのシステムイメージは，端末，ネットワーク，センターの 3 つの要素で構成されており，端末では，音声技術などを利用した知的な HMI（Human Machine Interface）が必須となっている．ネットワークはインターネットであり，使用者は端末から WEB 上の各種情報，サービスを受け取ることが可能である[5]．

2.2 車載応用での技術課題

車載応用では，安全性を確保するために音声技術は重要な技術である．しかし，現状ではまだ十分に仕様を満たしているとは言えない．特に，次の問題が存在している．

1)ユーザビリティの問題：

全てのインタフェース(HMI)は，透過的インタフェースの 3 原則を満足しなければならない．透過性インタフェースの 3 原則は，

①手順連想容易性の原則：システムをどう使えばよいかがすぐ分かる．

②状態理解容易性の原則：システムがどのような状態かがすぐ分かる．

③フィードバックの原則：利用者の行為の効果がすぐ分かる．

しかし，音声を利用したインタフェースでは，現状では，この透過性の原則を満足していない．入力音声が誤認識になると，状態が利用者の思考を越えた場面に飛躍して，結果として利用者は今の状況を理解することができない．音声認識技術の性能（認識率）が低いということにも関係している．

2)OOV(Out-Of-Vocabulary)の問題：

音声認識は語彙を設定しないと起動できない．結果として，種々の言い回しや短縮形に対しては語彙外発話になり，誤認識の原因となっている．前記のユーザビリティの問題とも大きく関係する．実世界の語彙を効率的に収集する方式や新規語彙，新規短縮語に関しての対処方法が課題となっている[6]．

3)頑強性の問題：

純粋に音声認識の問題である．車載応用では車環境の雑音(エンジン音，風切り音，路面ノイズ等)が大きな問題となっている[7]．自動車内での使用に際しては，エンジン音や空調等のノイズの問題や，マイクロフォンの位置が使用者の口元から離れている等の問題があり，認識率の低下が大きな問題となっている．音声認識では，発声者の音声が 30dB 程度のＳＮ比を持って入力されれば，ほぼ完全に認識が可能である．車載音声認識では，周囲ノイズ等の問題で，30dB のＳＮ比を確保するのが難しい環境となっている．走行状態でも，接話マイクロフォンならば，90%以上の認識率は確保できるが，サンバイザー等の離れた場所にマイクロフォンを設置した場合が問題となる．特に，窓あけで高速道路走行の状態では，40%程度の認識率しか得られていない．

3. 車載音声データと評価実験

3.1 車載音声（車載雑音）の特徴

車内音声認識を実現するためには，最も大きな問題は，車内雑音である．通常，100Hz 以下に存在する車のエンジン音のほか，走行時の雑音として，タイヤと路面との間で起こるタイヤ・路面雑音，窓を開けた場合の風きり音，ラジオ・カーステレオなどのオーディオ音，エアコン音，同乗者の声など様々な雑音が存在している．アイドリング時と市街地走行時の雑音スペクトルは，低周波成分が多く，特に，市街地走行時の低周波数成分が大きい．これは，車のエンジンノイズである．高速道路走行時の雑音スペクトルも同様に，低周波成分が大きい特徴があるが，タイヤと路面の間のノイズが高域周波数へも分布している．扱いにくいのは，タイヤの種類と路面の舗装材質の違いに大きく影響されて，スペクトル分布が大きく異なることにある．

3.2 車載音声認識のための雑音除去

今まで，検討されている車内雑音対応の音響信号処理に関して，簡単に概観する．

(a)低周波成分の除去

エンジンノイズ等の低周波成分の除去は，高域通過フィルター（HPF: High Pass Filter）で行うのが一般的である．カットオフ周波数は，50Hz～200Hz を想定している．この結果，少なくても 15dB 以上のＳＮ比を稼ぐことが可能である．

図 1 に，車載音声の状況を示す．図 1(a)は，静かな環境で発声された音声/詳細

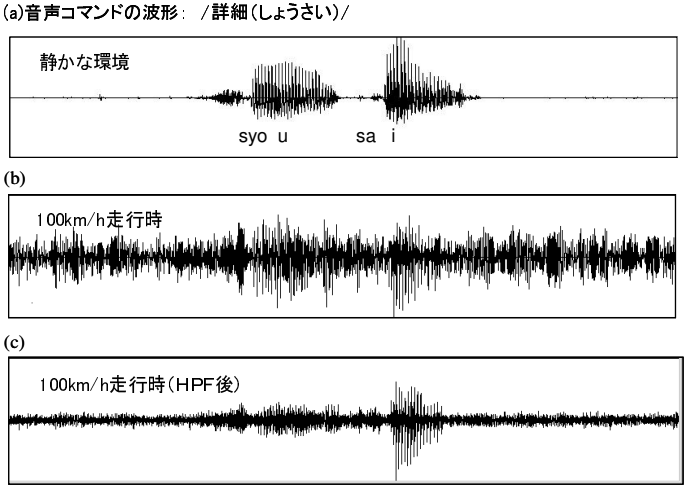

図 1　車載音声波形の状況

ⓒ 2010 Information Processing Society of Japan

（しょうさい）/の波形を示した．図 1(b)は，同じ/詳細/という言葉を，自動車内で発声した場合の波形である．自動車内の雑音により，音声波形が完全に隠れている（崩れている）．この波形に，カットオフ周波数 50Hz のＨＰＦ処理を行った波形を図 1(c)に示した．音声波形/詳細/が現れている．

(b)定常雑音成分の除去（スペクトルサブトラクション）

定常雑音の除去する簡易的な手法として，定常雑音のスペクトルを原音声スペクトルから減算する手法（スペクトルサブトラクション:Spectral Subtraction）がある．定常雑音の推定と，どのくらい減算をするかのノウハウがある[8]．

図 2 に，図 1(c)の音声波形にスペクトラムサブトラクションを施した場合の例を示した．まず，音声が発声される前の時間帯で，雑音を推定し，そのスペクトルを，原音声データ（音声＋雑音）のスペクトルから減算して，真の音声スペクトルを求めている．

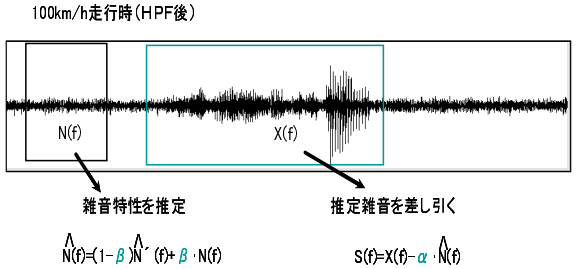

図 2　スペクトルサブトラクション処理

(c)アレイマイクロフォンの利用

自動車内では，発声者の位置が運転手の位置と固定することができるので，指向性の高いアレイマイクロフォンを利用すれば，周囲雑音を除去して，発声者の音声を取り出すことが可能である．遅延和アレーと減算アレーを組み合わせたビームフォーム(ABF: Adaptive Beam former)がある[9]．さらに，教師なし適応技術に基づくブラインド音源分離（BSS: Blind Source Separation）が，奈良先端大学院大学（猿渡等）から提案されている[10]．ABF の問題は，適応処理に伴う演算量の増加や，フィルター係数学習のために精度の良い教師情報（目的音声の無音区間と方位）を事前に必要とすることである．後者の BSS は，非常に自由度が高くプラグインで使用できる等の長所があるが,フィルター係数学習の難しさや残響による性能低下という問題がある.また，ビームフォームとスペクトラムサブトラクションとを組み合わせた手法が，北陸先端大学院大学（赤木等）から提案されている[11]．アレイマイクロフォン処理が，遅延和アレーと減算アレーのように，線形的な処理であれば，音響モデルの適応は不要であるが，通常は，音響モデルの学習のし直しや各アレイマイクロフォンの性能差の補正等が大きな問題であり，さらにコストの面からの問題もあり，まだ車載システムでの実用は少ない．しかし，確実に，発声者の音声を取り出せることがアレイマイクロフォン，3 チャネルでのアレイマイクロフォン応用は実現されると考えられる．

(d)音響モデルでの対応

(i)音響モデルの適応

自動車走行時での雑音を推定して，音声認識の音響モデルを適応化する手法が提案されている．しかし，適応した結果，確実に性能が向上することにはならず，適応化する場合の判断基準等の研究開発が今後の課題としてある．さらに，雑音モデルと通常の音響モデルとを合成するＨMM合成での対応も試みられている[12]．

(ii)雑音混入音響モデルの学習

車載雑音が推定でき，限定されていれば，雑音混入した音響モデルを学習することが一番効果がある．さらに，複数の環境での音声データをもとに，複数の雑音混入音響モデルを学習して，選択的に使用する方式もある．

3.3　車載音声認識用実験データ（収録環境）

今回使用した音声データは，（株）日立製作所中央研究所から入手した[1]．この音声データは，以下の条件のもとカーナビが使用される環境で収録された．

(1)自動車内に設置された遠隔マイクで，実際に走行中の音声
(2)単一マイクではなく，マイクロフォンアレイを利用
(3)孤立単語音声認識を対象
(4)読み上げ音声ではなく，自然発話に近い音声
(5)音声認識操作失敗時のデータをそのまま残す

特に(2)について，図 3 で示すようにマイクは助手席の前，ダッシュボード上に直線状に 7 つ配置し，それらの間隔は右から順に 10cm，5cm，5cm，5cm，5cm，10cm となっている．マイク番号も右から順に 1～7 となっている．平行して発話者に接話マイクを付け，これをマイク番号 8 とする．音声データは，都心部を走行して収録し，男女 5 名ずつ計 10 名 1967 発話である．POI（Point Of Interest）数は 152 個あるため，Julian において 152 個の POI 名を単語リストとして設定する．

表 1 に各マイクロフォンの SN 比と Julian での音声認識率 ASR を示した．SN 比は，音声の平均パワーS と無音部の平均パワーN を求め，SN=10logS/N[dB]で求めた．発声者の口元に近いマイクロフォン＃5 の SN 比が一番良く，次に＃4，＃3 と続き，両脇になるに従い，SN 比は悪くなっている．認識率（ASR 率）は，SN 比との相関がある結果となっている．しかし，マイク＃7 は，マイク＃2 と＃1 に比べて，SN 比は良いのに，認識率は劣化している.接話マイクロフォンで収録された音声の SN 比は 27.93 であり，認識率は 95.17%であった．この認識率が雑音除去処理の目標値となる．

ⓒ 2010 Information Processing Society of Japan

Vol.2010-SLP-82 No.4
2010/7/23

図 3　マイクの配置

表 1　各マイクロフォンから収録した音声の特性（SN 比と ASR 率）

	SN ratio [dB]	ASR rate [%]
Microphone #1	-3.23	83.22
Microphone #2	-2.71	86.27
Microphone #3	-0.96	86.83
Microphone #4	-0.32	87.80
Microphone #5	0.66	88.71
Microphone #6	-1.45	85.56
Microphone #7	-2.36	76.41
Headset microphone	27.93	95.17

3.4 大語彙連続音声認識ソフトウェア Julian

今回認識実験を行うにあたり Julian を使用した．これは HP 上で公開されており，無償で入手が可能である[13]．Julian は，有限状態文法(FSG)に基づく連続音声認識パーザである．Julian は言語制約以外のほとんどの部分を Julius と共有している．Julius は n-グラムで次の単語の予測を行っていくが，Julian はあらかじめ文法を設定する必要がある．今回は小語彙単語認識だけに絞って考えるため，認識率向上という観点から Julian を使用する．

4. 評価実験結果と考察

4.1 HPF 処理の効果の確認

カットオフ周波数 20Hz(男声)と 40Hz（女声）での HPF（High Pass Filter）を施した時の各マイクロフォンと各話者の認識率を図 4，図 5 に示した．さらに，マイクロフォン 5 の音声にカットオフ周波数を 20Hz から 100Hz に変異させた時の認識率の推移を図 6 に示した．図 4 と図 5 から，HPF の効果は，あまり見られないという結果となった．また，カットオフ周波数は，男声では 20Hz，女声では 40Hz の場合が，一番認識率が良いという結果になった．HPF による雑音除去の効果は，理論上ではあるはずなのだが，今回の実験においては大きな効果は観測されなかった．

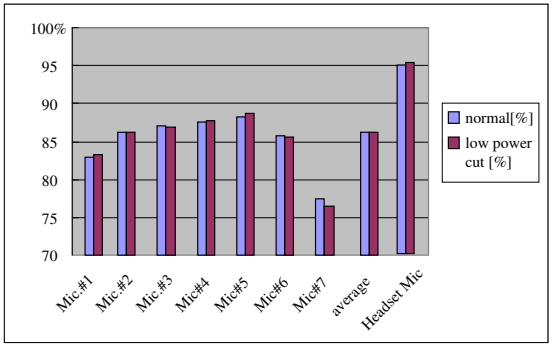

図 4　Low Power カットの効果（各マイクロフォン）

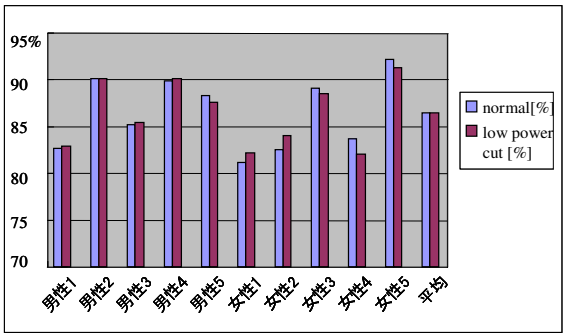

図 5　Low Power カットの効果（各話者）

ⓒ 2010 Information Processing Society of Japan

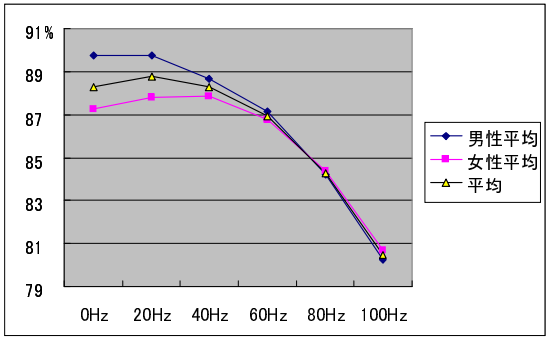

図 6　Low Power カットの効果（カットオフ周波数）

4.2　スペクトルサブトラクション（SS）とアレイマイクロフォン加重の個別効果

Linux マシン上で音声データを Julian にかけ，下記の手順で認識率を出した．

(1)発話者に近いマイク 4 のデータで男女 5 名ずつ計 10 名，50 単語，計 500 発話の認識率を算出する．

(2)スペクトルサブトラクション（定常雑音成分の除去（以後 SS））処理を行い，認識率を算出した．SS とは，図 2 で示すように音声が発声される前の時間帯で，雑音 N(f)を推定し，そのスペクトルを，原音声データ X(f)（音声＋雑音）のスペクトルから減算して，真の音声スペクトル S(f)を求める方法である．今回は発話者に近いマイク 4 のデータで男女 5 名ずつ計 10 名，50 単語，計 500 発話の認識率を算出した．また，音声が発声される前の時間帯の長さ（無音部）を適応的に変更させた．SS のパラメータは，Julian でのデフォルトである α=0.2，β=0.5 とした．

(3)さらに，アレイマイクロフォンによる音声の加重合成を行い，認識率を算出した．発話者に近いマイク 4 のデータを中心に．

①マイク 3＋マイク 4
②マイク 4＋マイク 5
③マイク 3＋マイク 4＋マイク 5

を加算し，男女 5 名ずつ計 10 名，30 単語，計 300 発話の認識率を算出した．

図 7 に SS 処理の効果を示す．その結果，10 名 50 単語平均で，SS 処理を行わない場合のマイク 4 の平均が 89.8%から，SS 処理を施した場合 92.6%と認識率の向上が見られ，接話マイクの平均 98.6%に近づく結果となった．次に，アレイマイクロフォン加重の効果を評価したが，大きな認識率の向上は図れなかった．個別評価結果は図 10 に示してある．

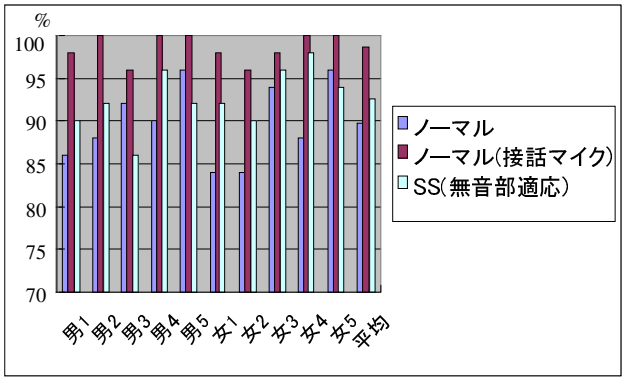

図 7　SS の効果（話者毎の認識率)

4.3　SS とアレイマイクロフォン加重の融合

アレイマイクロフォンは発話者に近いマイク 4 を中心に，4.2 で述べた 3 通りの加算を行い，さらに単語数を増やして各話者 100 単語の認識率を下記の手順で評価した．

(1)アレイマイクロフォンとスペクトルサブトラクション（以後 SS）を融合して認識率を算出する．

(2)Julian に加算した音声データを取り込み，下記の式で与えられる SS パラメータの α 係数の値を 2.0，3.0，3.5，4.0，5.0 と変え SS を行い，認識率を算出して比較する．

$$S(f) = X(f) - \alpha \cdot \hat{N}(f)$$
$$\hat{N}(f) = (1-\beta)\,\hat{N}'(f) + \beta \cdot N(f)$$
$$\hat{N}(f)\text{: estimated noise}$$

ここで，S(f)，X(f)，N(f)は，それぞれ真の音声スペクトル，観測された音声スペクトル，観測された雑音スペクトルを表している．α はフロア係数であり，減算した結果，音声スペクトルが負にならないようにするパラメータである．β は推定された雑音と過去の値とのスムージングを行うパラメータである．

(3)認識率第 2 位を算出し，第 1 位の結果と比較する．

図 8 にフロア係数 α を変化させた時の認識率の推移を示した．β 係数の最適な値は

ⓒ 2010 Information Processing Society of Japan

Julian でのデフォルト値の 0.5 であった．図には第 1 位候補と第 2 位候補の二通りの結果を示した．結果は，α=4.0 が最も大きく他のα係数よりも安定した結果になった．また認識率の高いα＝3.5，4.0 を使い認識率第 2 位を算出したところ，どのマイクも 1～2%認識率が上昇した．

図 9 に，アレイマイクロフォンと SS を融合したデータの認識率結果を示した．α=4.0 を使った．接話マイクには及ばないが，今回試みた処理の中でアレイマイクロフォン＋SS は一番良い認識率を出した．

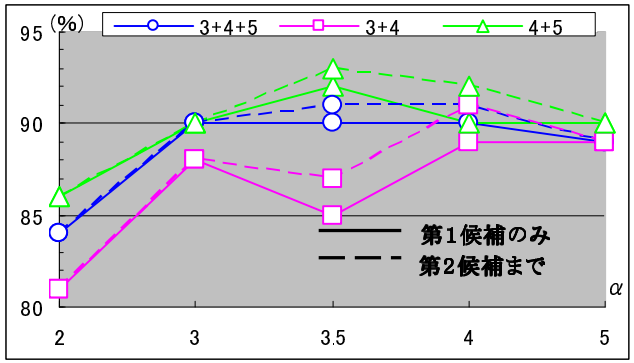

図 8　α係数と認識率の変化

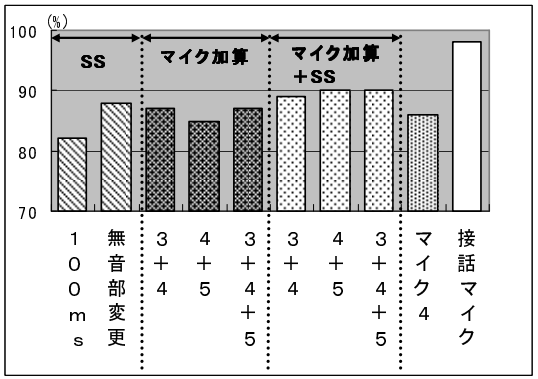

図 9　処理ごとの認識率の比較

5. おわりに

車載音声の雑音除去という課題で，SS とアレイマイクロフォン加重の融合を行った．その結果，種々のパラメータを適応化することで認識率の向上が図れた．今後の展開として，減算型アレイマイクロフォンによる雑音推定方式を検討して認識率のさらなる向上を目指したい．

謝辞　本研究で使用した車載音声データは(株)日立製作所中央研究所から入手した．経産省音声基盤技術プロジェクトで収集したデータである．大淵康成氏に感謝する．

参考文献

[1]大淵康成他，早稲田大学 IT 研究機構 音声技術実用化研究所「音声認識技術実用化に向けた先導研究成果報告書」C-3～C-52（平成 18 年 3 月）．
[2]高橋正幸，大瀧良一，畑岡信夫：「カーナビ音声認識の実用化に向けた車載音声の解析」，平成 20 年東北地区若手研究者発表会 pp.21-22(平成 20 年 2 月)．
[3]齋藤康夫，高橋英徳，畑岡信夫：「車載音声の解析と評価」，平成21年東北地区若手研究者発表会(平成21年2月)．
[4]上野聡，畑岡信夫：「車載音声の解析と評価(2) -アレイマイクロフォンとスペクトルサブトラクションの融合-」，信学会総合大会ISS学生ポスター（平成22年3月）．
[5]N. Hataoka, et al., “*Robust Speech Dialog Interface for Car Telematics Service*,” Proc. of IEEE CCNC2004 (Jan. 2004).
[6]K. Vertanen, “*Combining Open Vocabulary Recognition and Word Confusion Networks*,” ICASSP2008, SPE-P4, G7, Las Vegas (Apr. 2008).
[7]Y. Obuchi, et al. “Development and Evaluation of Speech Database in Automotive Environments for Practical Speech Recognition Systems,” Proc. of Interspeech2006, Pittsburgh, PA, USA (Sept. 2006).
[8]小窪浩明，天野明雄，畑岡信夫：「車載用音声認識における騒音対策とその評価」，電子情報通信論文誌D-II, Vol.J83-D-II, No.11, pp.2190-2197 (2000.11).
[9] Y. Kaneda et al., IEEE Trans. on ASSP, Vol.34, No.6, pp.1391-1400(1986).
[10] H. Saruwatari et al., EURASIP J. Applied Signal Processing, Vol.2003, No.11, pp.1135-1146 (2003).
[11] J. Li and M. Akagi, “A Hybrid Microphone Array Post-Filter in a Diffuse Noise Field,” Interspeech2005, No. 4CP2-8 (2005).
[12] F. Martin et al, “Recognition of Noisy Speech by Composition of Speech and Noise,” Proc. of European Conf. on Speech Communication and Technology, pp.1031-1034 (1993).
[13]大語彙連続音声認識エンジン Julius&Julian： http://julius.sourceforge.jp/.

ⓒ 2010 Information Processing Society of Japan