様式 C-19

科学研究費補助金研究成果報告書

平成　22 年　5 月　1 日現在

研究種目：　基盤研究(B)
研究期間：　2007　～　2009
課題番号：　19300060
研究課題名（和文）　実世界劣化音声コーパスに基づく音声強調法の研究
研究課題名（英文）　Study on Speech Enhancement Based on Distorted Speech Corpora in the Real-world

研究代表者
武田　一哉（TAKEDA　KAZUYA）
名古屋大学・大学院情報科学研究科・教授
研究者番号：20273295

研究成果の概要（和文）：実世界で雑音などにより劣化した音声の認識を目指し以下のことを行った。(1)劣化音声コーパスを整備し、CENSREC という名称で一般に利用可能とした、(2)劣化音声の認識率への影響度を測る指標を検討し、加法性・乗法性雑音に対して高精度に認識性能を予測できた、(3)劣化音声の劣化要因とその認識手法を体系化した、(4)劣化音声の認識手法を研究した。

研究成果の概要（英文）：For distorted speech recognition under the real world, we conducted below: (1) development of distorted speech corpora named CENSREC and distribution of them in public; (2) accurate recognition performance prediction for additively/convolutionally distorted speech; (3) development of structural explanation of distortion factors and recognition methods for distorted speech; (4) development of distorted speech recognition methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	6,200,000	1,860,000	8,060,000
2008 年度	4,700,005	1,410,000	6,110,005
2009 年度	3,700,000	1,110,000	4,810,000
年度			
年度			
総　計	14,600,005	4,380,000	18,980,005

研究分野：総合領域
科研費の分科・細目：情報学　・　知覚情報処理・知能ロボティクス
キーワード：劣化音声認識、劣化指標、対劣化手法体系化、音声認識評価基盤

１．研究開始当初の背景

本研究のメンバーは「情報処理学会・音声言語情報処理研究会」のもとで「雑音下音声認識評価ワーキンググループ」として、雑音下の音声認識を評価するために大規模なデータベースを収集し、それを利用した共通評価基盤の構築と配布を行うと同時に、雑音下音声認識の研究を行ってきた。その中で、現在の統計的な雑音抑圧手法は，音声と雑音の性質がコンパクトに表現可能であるという前提に立脚している。このため雑音と音声信号を事後的に計算機上で加算したような開発・評価データに対しては高い性能を示すが、これは劣化音声の空間を離散的にサンプリ

ングした有限の組み合わせを論じているに過ぎず、実環境下で頑健に動作する音声強調手法は劣化音声のブロードな分布に立脚することが必要であると考えた。

２．研究の目的

本研究は実世界劣化音声コーパスに基づく音声強調処理手法の確立が目的である。
雑音や残響が存在する日常の音環境の下で収録された音声を効率的に被服する実世界劣化音声データベースを設計・収集し，コーパスから抽出された劣化音声の分布を利用して，多様な要因により劣化した音声の品質を回復する方法を確立する。

３．研究の方法

次のような部分問題として本課題に取り組んだ。

(1) 実世界劣化音声コーパスの構築
大規模実世界コーパスを構築し、日本および世界で使用される標準基盤とする。
(2) 劣化指標の作成
雑音の劣化を数値化し認識率予測などに役立てる。
(3) 雑音・残響抑圧手法・対策手法の体系化
劣化音声やその対策手法を体系化し、コーパス構築や手法研究の指針とする
(4) 実世界分布での音声処理手法
構築した劣化音声を用いて劣化音声認識手法を研究する。

４．研究成果

(1) 実世界劣化音声コーパスの構築と配布

従来，雑音下音声認識のための手法が数多く提案されてきたが，これらの手法の性能を比較することは容易ではなかった．そこで，種々の手法を客観的に比較評価し，また競争を促すための共通評価環境が必要不可欠である．ただし、当初から非常に難しい問題に取り組んでも，効果的な研究開発は望めない。そこで、雑音などによる劣化要因およびそれらに対するアプローチを分類・整理し，個別要因を取り上げて順次コーパスを構築することとした．以下にそれらの成果物を列挙する．最後のロンバード音声 DB を除き，標準認識スクリプトや評価法まで規定した評価基盤である．

① CENSREC-1(AURORA-2J)

欧州の標準化機関 ETSI による加法性雑音シミュレーションの音声認識評価データベースを邦訳して作成された．連続数字発声に，計算機上で 8 種類の雑音を様々な信号対雑音比（SNR）で重畳した音声データベースを用いる．これまでに 100 部以上を配布している．ユーザは独自の雑音抑圧手法で雑音を抑圧したデータを作成すれば，音響モデルの学習から認識実験までが容易に行えるパッケージである．

② CENSREC-2

自動車内実環境下録音の連続数字認識評価環境である．CENSREC-1 との比較により，シミュレーションと実環境下の違いが評価可能である．

③ CENSREC-3

自動車内実環境下録音の孤立単語認識評価環境である．CENSREC-2 との比較により，連続音声と孤立単語という認識タスクの違いが評価可能である．

④ CENSREC-1-C

雑音下の音声認識の効果的な対処法として，音声/非音声区間の識別(Voice Activity Detection; VAD)を事前に行う方法が注目されている．本基盤は，認識の前処理としての VAD 性能を評価するものである．シミュレーション音声および実環境録音音声がある．この中で，音声認識に向けた新しい VAD 評価指標の提案も行った．

⑤ CENSREC-4

音声の劣化要因として，残響も無視できない．そこで種々の環境で収録したインパルス応答を CENSREC-1 と同様の連続数字に畳み込んだシミュレーション残響音声と，同じ環境で実際に発声して録音した実音声を認識評価できる基盤を構築した．

⑥ CENSREC-AV

音声が劣化した場合でもそれを発声する口の映像は影響を受けない．そこで口の動きの映像と音声との両方を用いてハイブリッドで認識を行う Audio-Visual 法も注目されている．本基盤は音声と同時に映像も収録し，その唇映像と，音声に雑音を重畳したものを同期させたデータに基づく評価基盤である．

⑦ ロンバード音声DB

CENSREC-1 と 2 に生じる差の要因として，雑音環境下で発声自体に変化を生じさせるロンバード効果がある．そのロンバード効果のみの影響を見るため，ヘッドフォンから雑音を聞きながら発声して収録したデータベースを作成した．本 DB は，ヘッドフォンをすることによる発声変形を避けるためにオープンエアタイプを用いるなど，様々な工夫の下に作成された貴重なものである．

(2)雑音・残響劣化指標の作成

音声認識技術のインタフェースへの応用を考えた場合にハンズフリーでの利用が有望である．その際には，マイクロフォンに音声が到達するまでに加法性の雑音が近接マイクよりも対音声比で大きく加わり，残響も無視できない大きさになる．これらと音声認識性能の関係を定式化できれば、使用環境によ

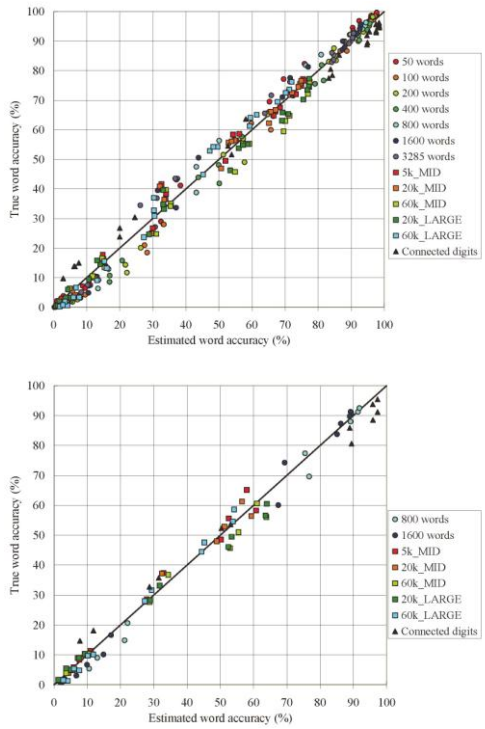

図 1：推定音声認識性能と実音声認識性能の対応(上：closed, 下：open)

って事前に認識性能が予測可能となり、対策手法を講じることが可能になる。

①□加法性雑音による音声認識性能予測
雑音環境下で、クリーンな音声に対して劣化した音声がどの程度のひずみが生じているかの評価値 x を求めることができた場合に、その値から認識率を予測することを考える。その際、認識タスクの難しさも考慮する必要があり、それを表す値 α も用いる。音声のひずみに対しては基本的に単語誤り率は単調増加することを踏まえて、シグモイド関数を用いて、次のような式を仮定する。

$$y = f(x,\alpha) = \frac{p_1\alpha^{q_1} + r_1}{1+\exp\left(-\left(p_2\alpha^{q_2}+r_2\right)\left(x-\left(p_3\alpha^{q_2}+r_2\right)\right)\right)}$$

ここで、α として SMR パープレキシティ(各単語の出現確率の逆数を単語パープレキシティとしたときの相加平均)を用いるのがよい。また、評価値 x としては PESQ を用いる。そして、式中の p_n、r_n、a_n を、様々なタスク(孤立単語認識、記述文法認識、大語彙連続認識を語彙数などで難易度をいくつか設定した計 13 タスク)の音声に雑音を重畳したものとその認識率から推定する。そして、こうして推定された式を用いて、PESQ と SMR パープレキシティから認識率を推定する。
推定に用いた音声で認識率を推定した結果(closed 条件)を図 1 上に示す。パラメータ推定データと同じデータであるので、本手法の上限と言ってよいが、実際の認識率と推定認識率のプロットが対角線近くに集まっており、非常によい対応がとれていることが分かる。さらに、推定に用いたものとは別に、実環境で収録した音声を用いて評価した結果(open 条件)を図 1 下に示す. パラメータ推定は雑音重畳によるシミュレーション音声を用い、実験音声は実環境収録であるが、それでも良い対応をしていることが分かり、シミュレーションで多くのパラメータ推定データを用意することで実環境の認識率推定が行えることが分かった。

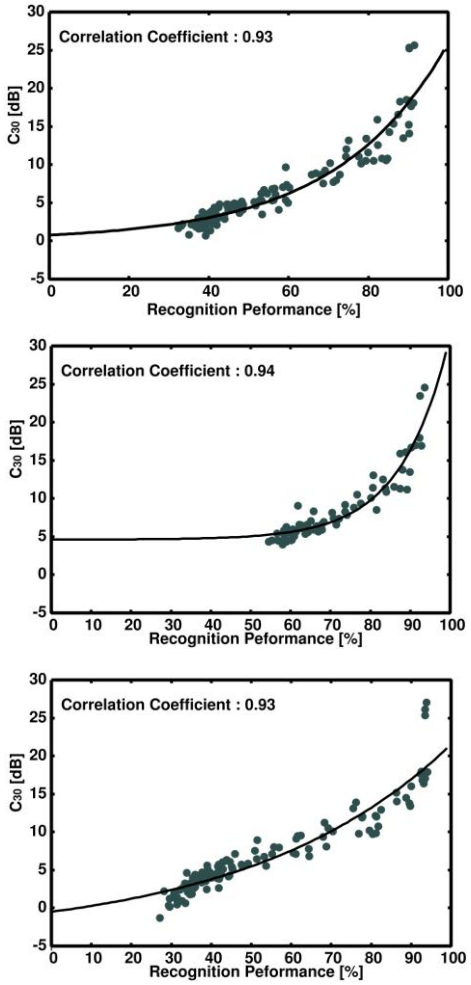

図 2：残響尺度からの認識性能予測曲線
(上から T_{60}=400, 600, 850 [ms])

②□残響下音声認識性能の推定
室内で残響を考える場合、音の初期部分の減衰状態を表現する指標である、室内音響指標が提案されている。このうち初期反射音と後続残響音のバランス指標の一つである C 値

$$C_{t_1} = 10\log_{10}\left(\int_0^{t_1} h^2(t)dt \Big/ \int_0^{\infty} h^2(t)dt\right)$$

すなわち、直接音と初期反射音のエネルギーに対する後続残響音のエネルギー比、中でも C_{30} が、音声認識性能と強い関連があることを見出した。さらに、各環境の C 値と音声認識性能の関係を指数関数を用いて回帰分析することによって，高精度に音声認識性能を推定できることを確認した．具体的な手法

内容として，まず予め音声認識性能を推定するための学習データを用いて音声認識性能と C 値に対して環境ごとに回帰分析を行う．そして算出した回帰曲線を残響時間ごとに分類後，推定用インパルス応答から算出した残響時間と C 値を基に算出した回帰曲線から音声認識性能の推定を行う．

本評価尺度を，前述した CENSREC-4 を用いて評価した．CENSREC-4 のうち、残響時間の異なる 3 環境(残響時間それぞれ T_{60}=400，600，850 [ms])の男女各 52 名(計 104 名)のそれぞれ連続数字音声計 4004 発話を用いて音声認識率および C 値を算出し、回帰曲線(指数関数を仮定)のパラメータを推定し，関係がそれぞれの回帰曲線で近似できているかを評価した。結果を図 2 に示す。各図左上に回帰曲線の相関係数を示している。

さらに、残りの 5 環境の認識率を C 値を用いて推定した。比較として従来の T_{60} を用いる推定も行った。結果を表 1 に示す。

表 1：音声認識性能推定結果

環境	推定尺度	認識率	推定認識率	誤差
A	T_{60}	93.1	70.5	**22.6**
	C値		92.6	**0.5**
B	T_{60}	54.3	70.5	**16.2**
	C値		56.9	**2.9**
C	T_{60}	74.1	56.0	**18.1**
	C値		85.2	**11.1**
D	T_{60}	65.3	56.0	**9.3**
	C値		60.7	**4.6**
E	T_{60}	30.7	52.3	**21.6**
	C値		50.9	**20.2**

図 2 より、各環境の相関係数が 0.93 を超えており，音声認識性能と C_{30} の関係を高精度に近似できた．これらの残響尺度を用いて性能推定した結果，提案手法の平均推定誤差値が従来手法より小さかったことより性能推定における提案尺度の有効性を確認した．これは性能推定に残響時間のみを用いる従来手法では性能推定値が系に関係なく一意に決定する問題点を各系のインパルス応答によって変動する C 値を用いることで解消できたためであると考えられる．しかし、環境 E(エレベータホール)での推定誤差が約 20%であることから高残響環境下の性能推定が困難であった．これは同じ C 値でも音声認識性能が異なる系が多数存在するためであると考えられる．逆にいえば、CENSREC-4 には幅広い残響特性をもったインパルス応答が収録されていることが確認できた．また提案手法は従来手法よりも高精度に推定できたが，更なる推定精度向上のために同じ C 値でも正確な性能推定ができる補正尺度の検討が今後の研究課題であると考えられる

(3)劣化音声とその対策の体系化

音声認識の入力音声は、様々な外乱を受けてマイクロフォンに到達する。本研究を通してそれら外乱を体系化、また対処する方法という側面からも整理した。

まず、音声の生成からマイクロフォン到達までの外乱を、

・ヒューマンファクタ $c[t]$

・残響や電装系による乗法性雑音 $h[t]$

・別音源の音波混入である加法性雑音 $n[t]$

と分類する。この時、本来発声しようとした音声を $s[t]$と最終的に観測される音声 $x[t]$の間の関係は

$$x[t] = h[t] * \left(c[t] * s[t]\right) + n[t]$$

と表せる。これらは音声に与える影響が比較的分離され、それぞれにアプローチがある。

・加法性雑音への対処

この対処法は、音声認識のステップのどこで実現するかにより、大きく 3 つに分類される。

1. 雑音に強い特徴量

最近多く研究される。一般的な特徴量では LPC<MFCC<PLP と言われる。また、特に対数スペクトルのコンポーネントごとの時系列にフィルタリングする手法が多く研究されている。

2. 雑音の抑圧

古くから研究されている。代表的なものにスペクトルサブトラクションやウィナーフィルタがある。また対数スペクトル領域で

$$X[t] = S[t] + H[t] + DCT\{\ln(1+\exp(IDCT(N[t]-H[t]-S[t])))\}$$

として第 1 項以外を劣化要因としてモデル化・追従・除去する方法がある。

3. 雑音対応音響モデル

雑音を付加した音声で音響モデルを学習する方法が基本である。それを未知の環境で実現するため適応手法や Parallel Model Combination(PMC)法を用いることもできる。

・乗法性雑音への対処

短時間の乗法性雑音に対しては、古くから用いられるケプストラム平均正規化法や分散まで正規化するケプストラム分散正規化法、あるいはケプストラムのヒストグラムを非線型に伸縮して学習時に合わせるヒストグラム正規化法が効果的である。しかし音声認識のフレーム長を超える残響などの伝達特性に対しては、かつては残響の逆フィルタリングなどを施して認識したりしたが、効果が低かった。最近では、遅れた残響成分は過去のフレームからの重畳成分とみなしてスペクトルサブトラクションの枠組みで正規化する方法などが用いられる。

・ヒューマンファクタ

音声認識は、決まり文句や書かれた文の読み上げのような丁寧な発話の認識で発展してきたが、機械を相手と意識しないような講演

や講義のような発話を認識することも必要となってきた。このような発話スタイルの違いは、発話の怠けや速度の変化、好い誤りや有声休止など様々な現象を含み、まだ対処法が確立された方法があるわけではないが、音響や言語モデルの適応などで性能が向上してきている。
また、雑音環境下でのロンバード効果もヒューマンファクタと呼べ、声が大きくなるとともに、スペクトルの第 1、第 2 フォルマントが高くなるシフト現象が観測され、それに対応したワーピング法が効果があるとされる。
・その他のアプローチ
1. VAD
音声以外の区間が無音の音響モデルより音声の音響モデルに適合して単語として認識される誤りが発生する。そこで，音声認識の前に，音の区間を事前に検出しておく(音声区間検出，VAD)。かつては有声音のパワーと無声子音部の零交差数への閾値処理から、実環境に対応するため、最近ではスペクトルや LPC 残差などの特徴量の高次統計量を利用した方法などが研究されている．また，音声と雑音のモデル化と尤度に基づいた識別手法も多く研究されている．これに、雑音のモデルを組み合わせて吸収させる方法を併用するのが一般的である。
2. マイクロフォンアレイを用いた方法
指向性マイクロフォンを用いる代わりに複数マイクをアレイ状に並べて信号処理で所望の音声を拾う方法が多数研究されている。
3. 映像情報を用いた方法
唇の映像が用いることができれば音声の劣化と直接の関係がないため、映像による読唇と音声認識を組み合わせる Audio-Visual 法も研究されている。
雑音下音声認識評価基盤の構築とそれに関する研究を行う経過で、このような体系化を行ってきた。それは研究基盤構築の裏付けにもなり、また研究推進方向の示唆にもなると考える。この結果は研究分担者による解説記事として日本音響学会誌に掲載された。

(4)劣化音声の認識手法
上記のような体系化の下、研究代表者および各研究分担者が様々な要因下の環境あるいはアプローチの劣化音声認識手法を研究した．詳細については紙面の都合上割愛するが、代表的なものを以下に挙げる。
・特徴量による加法性雑音の対処
ランダムプロジェクションによる音声特徴量抽出法
・加法性雑音抑圧による対処
反復スペクトルサブトラクションにおけるミュージカルノイズ低減法
確率モデルに基づく単一チャネル音源分離を用いた背景音楽抑圧
・長期残響の除去
マルチチャネル LMS アルゴリズムによるスペクトルサブトラクションと CMN に基づくブラインド残響除去
・VAD
VAD 信頼度による音響尤度補正に基づく VAD と音声認識の統合
・Audio-Visual 音声認識
マルチモーダル VAD によるマルチモーダル音声認識の精度向上

(5) 成果のまとめ
本研究では、劣化音声の認識という目的の下、本研究参加者のみならず日本や世界の研究者の研究の促進を図ることを考慮した。その結果、劣化音声コーパスを作成して国内外に配布し、多くの研究機関がこれを利用して研究を行うという成果を得た。また、このようなコーパスを作成するにあたり、劣化要因の影響度合いの計測の必要性から、特に加法性雑音と乗法性雑音について評価指標を考案し、音声認識率の高精度な予測を可能にした。さらに、コーパス作成のために、音声劣化要因やそれへの対処法を体系化することができた。そして、こうしたコーパスを用いて本研究参加者が多くの劣化音声認識手法を提案した。
このコーパスの作成や劣化指標の作成・体系化は国内外で評価を受け、今後のコーパスや指標などの規範となっていく可能性がある。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 11 件）

① 北岡教英, ″音声認識におけるロバストネス,″ 小特集---自動音声認識研究 の動向と展望---, 日本音響学会誌 Vol. 66, No. 1, pp. 23-27, 2010.(査読無)
② N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, S. Nakamura, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," Acoustical Science and Technology, Vol. 30, No. 5, pp. 363-371, 2009. (査読有)
③ N. Kitaoka et al., "Noisy speech recognition based on integration/selection of multiple noise suppression methods using noise GMMs," IEICE Trans. Inf. & Syst., Vol.E91-D, No 3, pp411-421,. 2008. (査読有)
④ 二宮芳樹, 宮島千代美(5 番目), 他. “音声と画像の統合によるドライバの発話区

間検出," 映像情報メディア学会誌, vol. 62, pp. 435-441, 2008. (査読有)
⑤ 原直, 宮島千代美, 伊藤克亘, 武田一哉. "多様な音響環境下における音声認識システム利用時のデータ収集システム," 信学論, vol. J90-D, pp. 2807-2816, 2007. (査読有)

〔学会発表〕(計 27 件)
① 小川哲司 他," ロンバード発声音声コーパスの設計と評価," 日本音響学会秋季研究発表会, 2009.9.15(郡山).
② 西川浩太郎, 森勢将雅, 西浦敬信, 南條浩輝, "反復スペクトルサブトラクションにおけるミュージカルノイズ低減法の検討," 日本音響学会秋季研究発表会, 2009.9.15(郡山).
③ 金正賢, 山田武志, 北脇信彦, "ETSI 標準雑音抑圧フロントエンドのための雑音推定法の検討," 電子情報通信学会 2009 年総合大会, 2009.3.19(愛媛).
④ 伊藤弘章, 西野隆典, 北岡教英, 武田一哉, "確率モデルに基づく単一チャネル音源分離を用いた背景音楽抑圧," 日本音響学会春季研究発表会, 2009.3.17(東京).
⑤ S. Tamura, C. Miyajima, N. Kitaoka, S. Hayamizu, K. Takeda, "CENSREC-AV: Evaluation frameworks for audio-visual speech recognition," Proc. AVSP, 2008.9.27 (Moreton Island, Australia).
⑥ M. Nakayama, T. Nishiura, Y. Denda, N. Kitaoka, K. Yamamoto, T. Yamada, S. Tsuge, C. Miyajima, M. Fujimoto, T. Takiguchi, S. Tamura, T. Ogawa, S. Matsuda, S. Kuroiwa, K. Takeda, S. Nakamura, "CENSREC-4: Development of evaluation framework for distant-talking speech recognition under reverberant environments," Proc. INTERSPEECH, 2008.9.24 (Antwerp, Belgium).
⑦ T. Nishiura, Y. Hirano, Y. Denda, M. Nakayama, "Investigations into early and late reflections on distant-talking speech recognition toward suitable reverberation criteria," Proc. INTERSPEECH, 2007.8.29(Antwerp, Belgium).

〔図書〕(計 1 件)
① COMPUTER PROCESSING OF ASIAN SPOKEN LANGUAGES, Consideration Books, c/o The American Group ,Mar. 2010 (4 Chapters by Takeda, Kitaoka, Tamura etc., 5 pages each, total 20 pages)

６．研究組織
(1)研究代表者
武田　一哉（TAKEDA KAZUYA）
名古屋大学・大学院情報科学研究科・教授
研究者番号：20273295

(2)研究分担者
北岡　教英（KITAOKA NORIHIDE）
名古屋大学・大学院情報科学研究科・准教授
研究者番号：10333501

山田　武志（YAMADA TAKESHI）
筑波大学・大学院システム情報工学研究科・准教授
研究者番号：20312829

西浦　敬信（NISHIURA TAKANOBU）
立命館大学・情報理工学部・准教授
研究者番号：70343275

宮島　千代美（MIYAJIMA CHIYOMI）
名古屋大学・大学院情報科学研究科・助教
研究者番号 90335092

田村　哲嗣（TAMURA SATOSHI）
岐阜大学・工学部・助教
研究者番号：10402215

(3)連携研究者
中村　哲（NAKAMURA SATOSHI）
独立行政法人情報通信機構・上席研究員
研究者番号：30263429

黒岩　眞吾（KUROIWA SHINGO）
千葉大学・大学院融合科学研究科・教授
研究者番号：20333510

柘植　覚（TSUGE SATORU）
徳島大学・大学院ソシオテクノサイエンス研究部・講師
研究者番号：00325250

滝口　哲也（TAKIGUCHI TETSUYA）
神戸大学・都市安全研究センター・講師
研究者番号：403978155

山本　一公（YAMAMOTO KAZUMASA）
豊橋技術科学大学・工学部・助教
研究者番号：40324230

小川　哲司（OGAWA TETSUJI）
早稲田大学・高等研究所・助教
研究者番号：70386598

中山　雅人（NAKAYAMA MASATO）
近畿大学・生物理工学部・講師
研究者番号：90511056