対話音声認識における環境や話し方の影響評定を備えた音声理解システムの研究

様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月３１日現在

機関番号：１３８０１
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：２１５００１６５
研究課題名（和文）　対話音声認識における環境や話し方の影響評定を備えた音声理解システムの研究
研究課題名（英文）　Study on a speech understanding system with an ability to estimate the effect of various environments and users on recognition accuracy
研究代表者
甲斐　充彦（KAI ATSUHIKO）
静岡大学・工学部・准教授
研究者番号：60283496

研究成果の概要（和文）：自動音声認識技術を利用する音声対話システムでは，周囲雑音や不明瞭な発話などによって随時起こり得る誤認識の可能性に対して適切に対処することが重要となる．本研究では，誤認識につながる諸要因に関わる特徴をユーザの発話から抽出し，認識性能を推測する技術の開発を進めると共に，対話状況に応じた認識性能の推定に基づいて音声対話システムで適切な応答選択を行う技術の開発を進め，評価実験において有効性を示した．

研究成果の概要（英文）： Designing a spoken dialogue interface involves an appropriate handling of recognition errors, which often caused by background noise or indistinct voice uttered by users, and has great impact on the usability of such interface system. This study has developed the methods for estimating user's recognition accuracy by his/her utterance, and also investigated the method to apply the recognition accuracy estimate, which depends on a dialogue state, for optimal selection of responses to the user. Evaluation experiments showed the effectiveness of the proposed methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,300,000	390,000	1,690,000
２０１０年度	1,100,000	330,000	1,430,000
２０１１年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

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：音声情報処理

１．研究開始当初の背景

自動音声認識技術を利用するシステムの実用化に関連して，普及の阻害要因として音声認識精度の低さだけでなく，課題の一つとして音声技術応用におけるシステム設計やユーザビリティの観点からの改善の必要性が指摘されている．例えば，音声インタフェースシステム（カーナビゲーションシステムなど）の使用中において期待した応答が得られなかったとき，音声だけでは目的（タスク）の解決が困難になることがある．そのような問題の解決のためには，音声認識・理解の正確さを随時そのときの対話状況や発話の特徴などに基づいてシステム側が自己判断し，ユーザに対して適切な応答を選択するような仕組みが望まれる．これまで，そのような仕組みは，システム開発者がインタフェース設計の段階で経験的に取り入れることが多

く，様々な環境や話者に対して有効な仕組みを実現するためにはその調整に多大な労力を必要としている．

２．研究の目的

これまで行ってきた音声認識・理解システムにおける発話検証や誤り率推定などの研究成果を基盤として，複数の視点から認識性能に関わる“聞き取り易さ”の情報を推定し，そのような情報に基づいて適切なフィードバックを選択する音声対話インタフェースシステムの実現へ向けて，基本的な課題を解決することを目的とする．具体的には，システムへの入力音声が音声認識・理解性能へ影響を与える特徴を“聞き取り易さ”と捉え，発話内容や話し方などに依存する発話単位の特徴や，雑音・残響などの外的な要因に依存する特徴などの観点から定量化して評定することを考える．また，そのような評定を音声対話システムにおける対話制御や応答生成などへ反映させるための枠組みの構築および評価を行う．

３．研究の方法

本研究では，先行研究での発話検証や対話音声理解に関する技術を，ユーザビリティ改善のアプローチである“聞き取り易さの情報のユーザへの直接または暗黙的フィードバック”を実現する手段として発展させ，様々な実環境を想定した音声対話インタフェースとしてのユーザビリティ改善と繋げるための基本的な課題を解決すること目的とする．そのようなシステムの構築および実証は，これまでに開発しているカーナビゲーションシステムを想定したタスクおよびそのタスク指向音声対話システムの要素技術を基盤として用いる．

具体的には，まず音声対話システムの使用中に音声認識・理解能力を知る手がかりを得るために，先行研究における発話検証や発音評定の方法を基本として，複数の視点で認識精度と関連した“聞き取り易さ”の指標として定量化することを試み，その実現方法と指標としての妥当性を実験的に明らかにする．このように推定する話者や外的要因の別による認識性能への影響の度合いは，クリーン音声の多数話者音声コーパスの話者適応音響モデル及び，代表的な雑音・残響特性の適応モデルを参照用モデルとしてあらかじめ用意し，それらのモデルとユーザの話者適応音響モデルとの間の比較による発音評定や誤り率予測に基づいて推定する．また，その結果を利用して現在の音声理解システムの音声理解手法や応答制御手法へ反映させるための枠組みの構築および評価を行う．そのために，雑音・残響を含む実環境を想定した被験者実験を行い，実環境や話者の違いの影響を含めて分析・評価を実施する．

４．研究成果

（１）複数の視点で認識精度と関連した“聞き取り易さ”の指標の推定に関して，音声認識システムで一般的なサブワード単位の統計的音響モデルから発話特徴を抽出する方法を基本として，予測モデルの構築および分析を進めた．まず，発話様式や話者の違いによる発話特徴の変化をよりよく捉えるため，発話速度(SR)，対数尤度(AL)，母音間距離(Dv)，同一音節間距離(Dd)，構造歪み(Ds)の５つの変量を定義した．発話速度以外の変量は，音声認識システムが用いる統計的音響モデルとしての不特定話者用 HMM と各話者の発話様式の特徴を反映した話者適応後の HMM から求める．変量の Dv,Dd,Ds の算出に用いる 2 つの音節モデル間の分布間距離は，音節単位 HMM の状態出力分布間距離として求める．同一音節間距離は，統計モデルによる音声認識スコアに最も関係する尺度として定義した．また比較するモデル間の構造歪みは，伝送歪みの影響を大きく受けず発話様式の違いや話者特有のなまりをおもに反映するという先行研究の知見に基づいて，発話特徴を表す別の観点として定義した．さらに母音間距離平均は，話者個々の音節分布が成す構造の大きさが明瞭度を反映すると想定して定義した．

まずモデル構築および評価の予備実験として，発話様式や話者の違いに焦点を当て，孤立単語を意図的に様々な発話様式で読み上げた音声を用いた．音声データは，ごく普通に孤立単語を読み上げた音声を 72 単語×3 発話(普通発話)，意図的に明瞭に孤立単語を読み上げた音声を 72 単語× 2 発話(明瞭発話)，それぞれ 7 名分使用した．7 名の音声データを用いた認識実験より全話者の平均認識率は普通発話に対して 87.90%，明瞭発話に対して 85.41%となった．明瞭に発話することが認識システムにおいて必ずしも良い結果に繋がるとはいえないことがわかる．各話者の発話様式の特徴の違いと認識率の関係を分析するため，目的変数を Pc(認識率)，説明変数を SR,AL,Dv,Dd, Ds として重回帰分析を行なった．まず，普通発話の音声データを用いた実験より話者毎の発話集合に対して各変量を求める．そしてそれらの変量を用いて以下の重回帰式の予測モデルを得た(各変量の値は正規化あり)．

$$Pc = -1.07Dd + 0.99Dv - 0.98Ds + 0.64AL + 0.31SR$$

この予測モデルにおける重相関係数は 0.87 で，話者単位での実際の認識率（約 77%～94%の範囲）に対する予測誤差（RMS）は 1.85%

となった．しかし，このモデルで発話様式が異なる明瞭発話に対しての予測誤差は 5.92%へ悪化することから，単語発話においても発話様式の違いによる認識性能への影響が大きいことが明らかとなった．つまり，タスク指向の音声対話システムのようにキーワード発話が多い想定であっても，発話様式の違いを考慮したモデル化が重要となることが示唆された．

更に，発話様式のバリエーションを 6 通りに増やした音声データを収録すると共に，雑音・残響の影響の違いによる影響分析を行うため車内環境での異なる走行状態での雑音・残響を含む音声データを用いて，上述のような発話特徴抽出の方法に基づく予測モデル化の実験および評価を行った．まず，発話様式のバリエーションを増やし，「通常」「早口」「遅口」「小声」「大声」「強調」の 6 種類で 9 名の単語発話を収録した．単語セットは，実環境車内単語音声データベース (CENSREC-3) にて定義されている 50 種類の孤立単語を用いた．また，語彙サイズが小さいタスクのため，音節正解精度として話者および発話様式の別での認識性能を評価し，前述の分析と同様に 5 種類の変量との関係を分析した．その結果，発話様式の違いによって相関が高い特徴量が異なる傾向がみられた．例えば，ほとんどの発話様式において，発話速度 (SR) と正解精度との相関係数はおよそ 0.55〜0.86 の範囲であったが，「大声」に対しては 0.25 と低くなった．また，話者の発話様式やなまりの違いを定量化することを意図した構造歪みでは，「早口」と「小声」ではそれぞれ相関係数が 0.8 と-0.57 と比較的高いが，それ以外の発話様式では小さかった．どの発話様式においても話者別の認識性能の違いを説明するのに有効な単独の特徴量はなく，単語発話という短い単位においても，話者によって発話様式の特徴に大きな違いをもたらしている可能性が示唆された．

続いて，雑音・残響の影響の違いによる影響分析を行うため車内環境での異なる走行状態の雑音・残響を含む音声データとして，実環境車内単語音声データベース (CENSREC-3) の男女計 17 名による孤立単語音声データを用いた．図 1 は，走行状態が「アイドリング」，「低速」，「高速」の別での話者別の連続音節認識での正解精度をプロットしたものである．連続音節認識では，新聞記事 4 年分 (約 330 万文) から学習した音節トライグラムの言語モデルを用いた．いずれの走行状態においても，話者別の認識性能の違いが大きく表れていることが分かる．

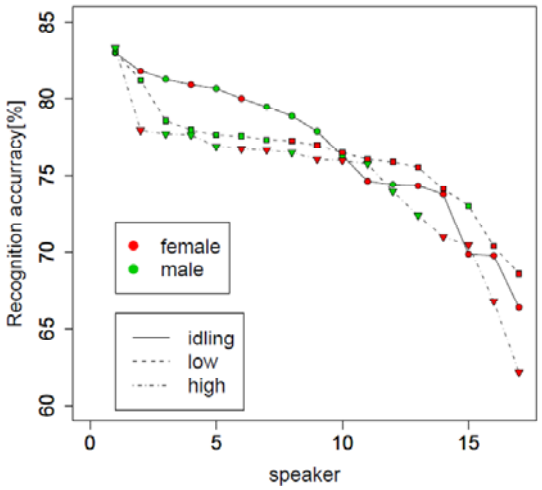


図 1 車内環境での走行状況別の話者別正解精度（男性 8 名，女性 9 名，正解精度順にソート）

また，前述のような発話特徴の抽出法によって，走行状態および話者の別に求めた特徴量および正解精度の相関係数行列を表 1 に示す．この結果より，走行状態によって雑音レベルの違いが大きいにも関わらず，雑音レベルを示す SNNR の特徴量と正解精度 (Acc) との相関が最も高いとはいえず，明瞭性の定量化を意図した母音間距離 (Dv) の方がやや相関が高いことが分かった．また，走行状態別に分析しても，母音間距離 (Dv) はほぼ一定してやや大きい相関係数 (0.5〜0.6) を持つのに対し，SNNR は 0.22〜0.57 の範囲で大きな変動があった．これらの結果から，話者毎の声の大きさと明瞭性などが相互に関係し，認識性能の違いに大きく影響している可能性があり，SNNR は算出が容易であるが予測材料として直接用いることが難しいことが示唆された．

	Acc	SR	SNNR	Dv	Dd	Ds	LL
Acc	1.00						
SR	-0.19	1.00					
SNNR	0.44	0.29	1.00				
Dv	0.53	0.33	0.68	1.00			
Dd	-0.36	-0.10	-0.22	-0.13	1.00		
Ds	0.11	-0.19	-0.10	-0.05	0.37	1.00	
LL	-0.14	-0.40	-0.65	-0.60	-0.17	0.20	1.00

表 1 発話特徴量と正解精度の相関行列

これまでの分析では話者数がやや少ないため，話者別の認識性能との関係の分析が不十分であった．そこで，多数話者を収録した JEIDA 日本語共通音声データベースから男女計 134 名分の単語発話を使用した分析を行った．各話者の発話内容は共通で，全国の地名に関する 100 単語の発話からなる．図 2 に連続音節認識での正解精度を話者別にソートした結果を示す．

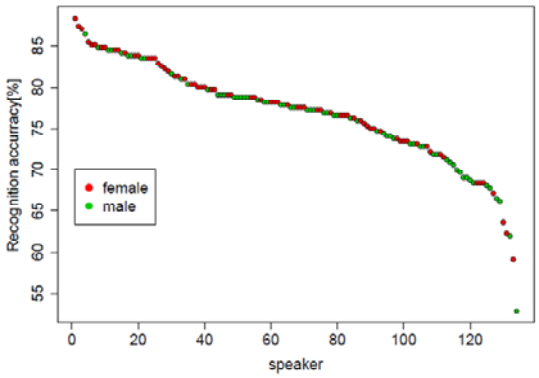


図 2 孤立単語発話の連続音節認識における話者別正解精度（男性 75 名，女性 75 名，正解精度順にソート）

この結果においても，認識性能が最も高い話者から低い話者まで，広い範囲に話者別の認識性能が分布していることが分かる．この多数話者データに対して，前述のように話者別での発話特徴および音節正解精度（Acc）について相関分析を行った結果を表 2 に示す．この結果においても，母音間距離（Dv）と正解精度との相関は他の特徴量と比べて比較的高く，次いで同一音節間距離（Dd）および構造歪み（Ds）となった．つまり，話者適応によって求めた統計的音響モデルから抽出した話者特徴が有効であることが示された．しかし，各話者毎のデータをもとに重回帰分析を行った結果，重相関係数は 0.34 であり，予測モデルとして不十分なものとなった．

表 2　多数話者データでの相関分析結果

	Acc	SR	SNNR	Dv	Dd	Ds	LL
Acc	1						
SR	0.09	1.00					
SNNR	0.14	-0.03	1.00				
Dv	0.45	0.03	0.30	1.00			
Dd	-0.25	0.15	-0.17	0.10	1.00		
Ds	-0.24	0.29	-0.09	-0.34	0.31	1.00	
LL	-0.01	-0.13	-0.24	-0.58	-0.43	0.10	1.00

そこで，より詳細に分析をするため分析データを Acc が 70%以上(C_{High})，70%以下(C_{Low})で 2 クラスに分け，クラス別に各特徴量の平均値，標準偏差を求めた．その結果，クラス間で相対的にみると，SNNR，Dv の値は，認識精度の上昇に伴い大きくなり，逆に Dd, Ds の値は小さくなっており，前の相関分析の結果で示されている各特徴量と認識精度間に存在する仮説に矛盾しない形となった．また，6 種類の特徴量の値の分布は，各特徴量の標準偏差は，低クラスの方が大きくなっており，認識精度の低い話者集団における話者間での特徴量のばらつきが認識精度の予測を難しくしている原因のひとつと考えられた．

上述のような分析結果を踏まえ，大まかな認識性能の違いを推定することが可能か確かめるため認識性能のクラス判別（2 クラス）の実験を試みた．クラス判別の方法としては，（1）重回帰モデル，（2）ロジスティック回帰モデル，（3）ベイジアンネットワーク，の 3 種類のモデルを構築して評価した．表 3 の交差検証による実験結果が示すように，重回帰モデルでは認識性能が特に低い話者をある程度検出可能であることが示された．他のモデルで判別性能が改善されなかった原因としては，話者単位のサンプルで学習するため，モデルパラメータの数に対して学習サンプル数（話者数）がまだ十分でないことが考えられる．

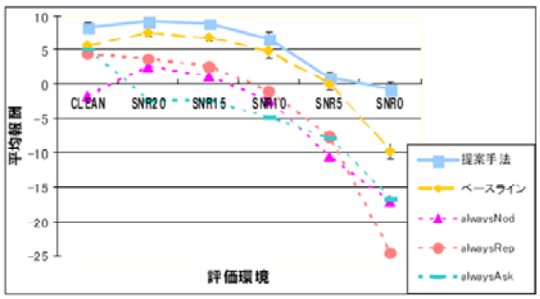
表 3　認識性能のクラス判別結果

	重回帰モデル		
Acc 境界 (%)	high	low	判別率 (%)
70	113/115	4/19	87.3
75	76/91	21/43	72.4
80	14/40	81/94	70.9
	ロジスティック回帰モデル		
Acc 境界 (%)	high	low	判別率 (%)
70	113/115	3/19	86.6
75	83/91	17/43	74.6
80	5/40	79/94	62.7
	ベイジアンネットワーク		
Acc 境界 (%)	high	low	判別率 (%)
70	86(25)/115	0(11)/19	64.2
75	75(6)/91	16(12)/43	67.9
80	11(9)/40	48(37)/94	44.0

（2）音声認識性能の推定に基づいて，音声理解システムの意図理解手法や応答制御手法へ反映させるための枠組みの構築および評価を行った．音声認識誤りの可能性を考慮した対話制御を設計するとき，対話中の各ターンの発話意図の認識誤りを考慮するだけでは十分ではなく，現在の対話状況の不確かさを適切に考慮して対話制御を行う必要がある．このような問題に対して，部分観測マルコフ決定過程（POMDP）のモデルを応用するアプローチが近年注目されている．POMDP では，現在の不確かな状況の信念分布に基づいて，将来得られる期待報酬を最大化する方策（行動）を強化学習によって最適化する．本研究では，音声対話システム（SDS）において最適な対話制御を与える問題を POMDP によって定式化し（SDS-POMDP），その中で観測モデルとしてのユーザ意図認識の確率モデルを，ユーザの発話意図毎に異なる認識性能の推定値として与え，最適な対話制御の問題を扱う方法を考えた．SDS-POMDP において，状態はユーザのゴール，対話履歴，ユーザアクション（意図）の 3 つ組としたが，小規模のスロットフィリング型のタスクにおいて

も状態数が膨大となり，強化学習で扱うのが困難となる．そこで，本研究では 2 つのスロット（県名および市名）からなる小規模の地名入力タスクを想定し，ユーザアクションの種類は県名のみ（7 種類），市名のみ（42 種類），県名と市名の同時発話（42 種類）など計 94 種類を定義した．また，対話制御に用いるシステム側からの応答の種類として，県名のみの確認，市名のみの確認，直前のユーザ発話の推定意図の復唱のみ，これまでの対話で埋まっているスロット内容の確認，など計 186 種類を定義した．この結果，SDS-POMDP として扱う全状態数は 42×94×6＝23,688 個となった．このような条件において，ユーザ意図別の認識誤りモデルを SDS-POMDP における観測モデルとして適用し，音声認識誤りを疑似的に与えた対話シミュレーションによって評価実験を行った．

図 3 の結果は，提案法と比較のため用いた 4 つの方法について 500 回の対話シミュレーションにおける性能を比較したものである．ここで，“ベースライン”は観測モデルにおいてユーザの発話意図の違いに関わらず一定の認識誤り率を仮定した手法，他の 3 つはそれぞれ“alwaysNod”は常に相槌，“alwaysRep”は常に復唱，“alwaysAsk”は常に情報要求，という単純方策を用いた手法である．また，報酬は対話ターン毎のシステムからの応答に対して定義されるもので，一つの対話が完了するまでのターン毎に-1，ユーザゴールと矛盾する応答をした場合に-5，最終的にユーザゴールを確定できた場合に+10 のように定義している．そのため，対話当たりの平均報酬の大きさは必要としたターン数の少なさや，誤った応答の少なさを反映している．図 3 の結果より，ユーザの意図毎の認識性能の推定を用いて学習した方策を用いる提案法は，どの雑音レベルの評価環境でも高い平均報酬を得ており，有効性が示された．



(b)環境B:AER=15.0%(CLEAN)～43.4%(SNR0)

図 3 応答生成の方策の違いと平均報酬（横軸は評価時の雑音レベルの違い，各折れ線は比較するそれぞれの方策を表す）

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　8 件）

① 赤尾佳彦，甲斐充彦，王 龍標，“話者や発話固有の特徴の違いに注目した認識性能の個人差の要因分析”，日本音響学会 2012 年春季研究発表会講演論文集，査読無し，2012，pp. 3-P-2.

② 西島祥悟，甲斐充彦，小暮悟，王龍標，“音声認識誤り率の推定を用いた POMDP モデルの構築の検討”，人工知能学会言語・音声理解と対話処理研究会資料，査読無し，2012，pp. 13-19.

③ 野末隆史，小暮 悟，甲斐充彦，小西達裕，伊東幸宏，“音声対話制御のための HIS-POMDP 学習・評価プロトタイプツールの開発”，人工知能学会言語・音声理解と対話処理研究会資料，査読無し，2012，pp. 21-26.

④ 王 龍標，岸 良樹，張　兆峰，甲斐充彦，“複数の人工室内インパルス応答を用いた残響モデルの利用による遠隔発話話者認識”，日本音響学会 2011 年秋季研究発表会講演論文集，査読無し，2012，pp. 2-10-6.

⑤ 張　用起，甲斐充彦，王　龍標，“単語断片の候補選択が可能な音声入力インタフェースの実装と評価”，情報処理学会研究報告，査読無し，Vol. 2011-SLP-89，No. 25，2011，pp. 1-8.

⑥ Yonggee Jang，Atsuhiko Kai and Longbiao Wang，“Multimodal Interface with N-best Display Including Candidates of Spoken Word Fragments”，Proceedings of 2nd. APSIPA Annual Summit and Conference，査読有り，2010，pp. 478-481.

⑦ 尾崎，小暮，甲斐，小西，伊東，“複数の車内機器操作と雑談を扱えるマルチタスク音声対話システムのユーザビリティの向上”，情報処理学会研究報告，査読無し，Vol. 2010-SLP-80，No. 6，2010，pp. 1-6.

⑧ Yonggee Jang，Atsuhiko Kai and Longbiao Wang，“Speech Interface for Isolated Words Based on Combination of Search Candidates from the Common Word Parts”，Proceedings of Western Pacific Acoustics Conference (WESPAC X 2009)，査読有り，2009，pp. 0261 (7 pages).

〔学会発表〕（計　8 件）

① 西島祥悟，甲斐充彦，小暮悟，王龍標，“音声認識誤り率の推定を用いた POMDP モデルの構築の検討”，人工知能学会言語・音声理解と対話処理研究会，

2012.3.26（東京大学）.

② 野末隆史,小暮 悟,甲斐充彦,小西達裕,伊東幸宏, “音声対話制御のためのHIS-POMDP 学習・評価プロトタイプツールの開発”, 人工知能学会言語・音声理解と対話処理研究会, 2012.3.26（東京大学）.

③ 赤尾佳彦, 甲斐充彦, 王 龍標, “話者や発話固有の特徴の違いに注目した認識性能の個人差の要因分析”, 日本音響学会2012年春季研究発表会, 2012.3.15(神奈川大学)

④ 王 龍標, 岸 良樹, 張 兆峰, 甲斐充彦, “複数の人工室内インパルス応答を用いた残響モデルの利用による遠隔発話話者認識”, 日本音響学会 2011 年秋季研究発表会, 2012.3.15（神奈川大学）.

⑤ 張 用起, 甲斐充彦, 王 龍標, “単語断片の候補選択が可能な音声入力インタフェースの実装と評価”, 情報処理学会音声言語情報処理研究会, 2011.12.20（芝浦工業大学）.

⑥ Yonggee Jang, Atsuhiko Kai and Longbiao Wang, “Multimodal Interface with N-best Display Including Candidates of Spoken Word Fragments”, 2nd. APSIPA Annual Summit and Conference, 2010.12.16（シンガポール・Biopolis）.

⑦ 尾崎, 小暮, 甲斐, 小西, 伊東, “複数の車内機器操作と雑談を扱えるマルチタスク音声対話システムのユーザビリティの向上”, 音声言語情報処理研究会, 2010.2.12（兵庫県・須磨温泉）.

⑧ Yonggee Jang, Atsuhiko Kai and Longbiao Wang, “Speech Interface for Isolated Words Based on Combination of Search Candidates from the Common Word Parts”, Western Pacific Acoustics Conference (WESPAC X 2009), 2009.9.21（中国・北京）.

６．研究組織

(1)研究代表者

甲斐 充彦（KAI ATSUHIKO）
静岡大学・工学部・准教授
研究者番号：60283496

(2)研究分担者

小暮 悟（KOGURE SATORU）
静岡大学・情報学部・講師
研究者番号：40359758

王 龍標（WANG LONGBIAO）
静岡大学・工学部・助教
研究者番号：30510458

(3)連携研究者

なし