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２８日現在

機関番号：１２６０８
研究種目：基盤研究（Ｂ）
研究期間：2009～2011
課題番号：２１３００００６３
研究課題名（和文）　個性及び表現性ロバストな音声言語インタフェースに関する研究

研究課題名（英文）　Research on robust spoken language interfaces for diverse voice variability and expressivity

研究代表者
小林　隆夫（KOBAYASHI TAKAO）
東京工業大学・大学院総合理工学研究科・教授
研究者番号：70153616

研究成果の概要（和文）：ユーザの嗜好や気分に応じた表現豊かな音声出力と，ユーザの個性，気分や話し方の変化に頑健な音声入力ができる音声インタラクションを実現するためのロバスト音声認識・合成技術の確立を目指して研究を行った。ロバスト音声合成では，基本周波数量子化に基づく韻律コンテキストや自然発話・会話音声合成のための拡張コンテキストに基づく音声合成手法を，またロバスト音声認識では，感情表現・発話様式などのパラ言語情報の検出・表出度合の推定手法及び高速なモデル適応手法を確立し，その有効性を示した。

研究成果の概要（英文）：The purpose of the research is to develop techniques that make the human-computer interaction using speech input/output more robust for variations of users' emotional states, speaking styles, preferences, and expressivity. We have proposed techniques using a quantized fundamental frequency prosodic context for robust speech synthesis and an extended context set for spontaneous conversational speech synthesis. We have also proposed techniques for robust speech recognition including extraction of paralinguistic information and rapid model adaptation.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	2,900,000	870,000	3,770,000
２０１０年度	2,700,000	810,000	3,510,000
２０１１年度	1,900,000	570,000	2,470,000
年度			
年度			
総　計	7,500,000	2,250,000	9,750,000

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

情報技術の高度化やユビキタス情報環境の普及は，我々の生活をより一層便利にする反面，高度な情報機器や情報環境を使いこなせるかどうかにより新たな社会的格差を生みだしている。一方で，情報化社会の恩恵を享受しているユーザにとっても，より個人の嗜好に合った，かゆいところに手が届くようなサービスへの要求が高まっている。

これらの問題の解決に向け，人間にとって自然で違和感のないインタフェースを持つ知的インタラクションシステム実現の研究が重要さを増しており，その基盤の一つである音声インタフェースにおいても，音声に含まれる言語情報のみを対象とするのではなく，感情（気分），発話様式，個性，意図，態度といったパラ言語・非言語情報の抽出及びその利用の研究が活発に行われている。

このような状況において，「気が利いて融通のきく音声インタラクションを実現するために必要な要素技術は十分か」と問われた場合，残念ながら現実にはまだ不十分と言わざるをえない。例えば音声認識においては，原稿読上げ音声に対して認識率 95%を越えるような高い性能が達成されているが，日常に現れる様々な感情や発話様式が含まれる音声，自然発話音声や会話音声が対象となると，いまだに十分な性能が得られていない。音声合成においても，限定された話者の特定のスタイルを持つ音声を出力する目的であれば，高品質の合成音声出力が可能となりつつあるが，任意の話者性，個性，多様な発話様式・感情の表現を自在に制御することは実用化されていない。人間同士のいわゆる「相手の顔色を窺いながら」柔軟に対応できるマルチモーダルインタフェースの実現は依然として困難な状況にある。

これに対し本研究では，これらの要求に応えるためのロバスト音声合成・認識に関する新たな要素技術の開発を行うことを意図している。

２．研究の目的

「個性及び表現性ロバストな音声インタフェース」とは，気が利いて融通のきく音声を用いた情報のやりとりにつながる概念であり，究極的には以下のような音声インタフェースの実現を想定している。

(1) ユーザの意図を理解し，ユーザの気分や好みを読み取ってくれる。
(2) ユーザの嗜好や気分に応じて自由にカスタマイズ，パーソナライズが可能。
(3) 音声入出力において，ユーザに不自然さや違和感を感じさせない。

本研究では，このような音声インタフェースの実現につながる基盤要素技術として，平均声とスタイル制御に基づくロバスト音声合成に関する２項目と，スタイル推定とスタイル適応に基づくロバスト音声認識に関する３項目からなる以下の５項目の課題について，新たな手法の提案とその有効性の検討を行うことを目的とする。

(1) 感情表現・発話様式・声質を制御可能な音声合成
(2) 自然発話・会話音声の合成
(3) パラ言語情報の検出・表出度合推定
(4) 話者・スタイル変動に頑健な音声認識
(5) 動作からのパラ言語情報の抽出

３．研究の方法

(1) 感情表現・発話様式・声質を制御可能な音声合成

所望のスタイル（音声の感情表現・発話様式），声質，話者性，個性などのパラ言語・非言語情報を含む表現力豊かで多様な音声の合成手法として，研究代表者はこれまでにスタイル制御と呼ぶ手法を提案し，発話様式や感情（気分）などを含むある一つの話し方の調子（スタイル）を一つの次元に対応させた低次元のパラメータを変化させることにより合成音声のスタイルを直観的に制御可能なこと，声質の制御にも応用可能なことを示した。本研究では，この手法において，学習データの偏りや韻律ラベルの不正確さに対してより頑健なモデル学習を可能とするために，新たな韻律コンテキストを導入し，詳細な検討を行う。また日本語以外の言語への適用や声質変換などへの応用を検討する。

(2) 自然発話・会話音声の合成

少量のモデル学習用音声のみを用いて自然な合成音声の生成を可能にするロバスト音声合成の枠組みを，任意話者の自然発話・会話音声合成に発展させる。本研究では，平均声方式に基づく新たな手法の検討に加え，日本語話し言葉音声コーパス（CSJ）を利用した新たな韻律コンテキストセットの導入や，学習データが少量であることに起因する問題を緩和するためのモデルパラメータ共有決定木の構築方法を検討することにより，自然な韻律を持つ対話音声合成を実現する。

(3) パラ言語情報の検出・表出度合推定

スタイル制御手法の逆過程問題を定式化することにより，重回帰隠れマルコフモデル(重回帰 HMM)に基づいて感情（気分）や発話様式など，音声に含まれる種々のスタイル情報を表す直観的なパラメータを推定する手法を確立する。さらに，このスタイル推定手法に基づくパラ言語の検出・表出度合推定システムを構築し，発話様式の識別への応用の検討を行う。

(4) 話者・スタイル変動に頑健な音声認識

前述の課題(3)で確立した音声のスタイル推定手法を利用し，多様な話者性の変動だけでなく，音声に含まれる感情表現・発話様式の変動に対して，オンラインで高速に適応する手法を提案する。これを，感情表現を含む音声認識や一般話者の自然発話・会話音声の認識に適用してその性能評価を行い，提案手法の有効性を明らかにする。

(5) 動作からのパラ言語情報の抽出

話し手のしぐさから相手がどのような気分で話しているかを推定する手法の開発のための基礎検討を行う。本研究では，会話中の顔の姿勢推定，より具体的には頷きや傾げに着目し，連携研究者と協力して，ビデオカメラにより撮影した顔画像を対象として，特徴量抽出手法の検討を行う。

４．研究成果

(1) 感情表現・発話様式・声質を制御可能な音声合成

最近の音声合成研究の主流である隠れマルコフモデルに基づく統計的パラメトリック音声合成(HMM 音声合成)の枠組みにおいて，基本周波数(F0)量子化に基づく韻律コンテキストを導入する新たな手法を提案した。

HMM 音声合成では，音素などの音韻単位毎に，隠れマルコフモデル(HMM)を用いてスペクトル，F0 及び音韻継続長特徴量をあらかじめモデル化しておく。そして，入力テキストが与えられた場合，その読みに対応した音韻単位 HMM を並べて接続することにより一発話に対応する HMM を構成し，これから音声合成に必要なパラメータ系列を生成している。ここで，より自然な合成音声を生成するために，音韻単位 HMM を一つの音韻に対して１個だけ用意するのではなく，前後の音韻の並び方，品詞情報，アクセントの有無，前後のアクセント句の並び方，発話中の位置など，音韻や韻律に影響を及ぼす可能性のある変動要因（コンテキスト）を考慮し，同じ音韻単位でもこれらのコンテキストが異なる場合は別個のモデル（コンテキスト依存モデル）として表現している。

コンテキスト依存モデルを作成する際に必須となるのが，モデル学習用の音声データとそれがどのようなコンテキスト環境にあるかを記述したコンテキストラベル系列であり，音声データの量や質と共に，含まれるコンテキストの豊富さと対応するラベル情報の正確さが，合成音声の自然性に大きく影響することが知られている。音韻，品詞，文中の位置などの言語情報は比較的精度よく自動ラベリングが可能である反面，アクセントの有無やアクセント位置などの韻律情報は個人性やスタイルによる変動が大きく，自動ラベリングの精度が十分に得られず，人手による修正が必要となる問題があった。

これに対し，提案した量子化 F0 コンテキストは，従来のアクセントの有無や位置をコンテキストとするのではなく，対数基本周波数(logF0)の相対的な概形を量子化し，有限個のシンボルで表現してコンテキストとしている。具体的には，絶対的な声の高さによる影響を取り除くために，まず発話中の logF0 の分布を正規化した後，音韻単位で平均値を求め，これを量子化してシンボルに変換している（図 1 参照)。このため，通常の音声分析により F0 が求められさえすれば，自動的かつ教師なし学習によりラベル付けを行うことができ，人手によらない頑健な韻律コンテキストラベル付与が可能となる。

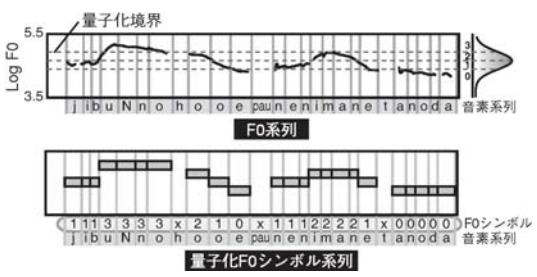

図 1 基本周波数(F0)パタンと量子化 F0 シンボル系列の概念図

平均声方式に基づく音声合成に提案コンテキストを適用し，量子化レベル数を変化させた場合の F0 歪との関係を調べた結果（図 2 参照)，量子化レベル数が 4 以上で客観評価値に大きな差がないことがわかり，主観評価でも量子化レベル数が 8 以上では差を知覚できないことが確かめられている〔雑誌論文⑧〕。また，提案コンテキストを声調言語（韻律が重要な言語情報の一部になっている）の一つであるタイ語に適用した〔学会発表⑥〕結果，学習用音声中の声調誤りや不明確な声調に対して頑健な韻律モデル学習が可能となり，合成音声の声調誤りによる不自然さを減少できることを明らかにした〔雑誌論文③〕。さらに，提案コンテキストは，声質変換〔雑誌論文④，⑤〕や極低ビットレート音声符号化〔雑誌論文②，学会発表⑤〕のためのモデル化の際にも有効であることを示した。この他にも，HMM 音声合成における韻律コンテキストの種類と日本語及び英語音声合成音声の自然性に関する詳細な検討〔雑誌論文⑥〕や，合成音声の話者個人性の強調手法〔学会発表⑧〕，不特定話者を対象とした音声のスタイル制御法〔学会発表②〕を提案した。

これらの成果は，音声資源が十分で整備されていない(under-resourced)言語における音声合成のための頑健なモデル学習や多様で表情豊かな音声合成システム実現に向けて有用になると考えられ，今後のさらなる研究展開や応用が期待できる。

(2) 自然発話・会話音声の合成

対話音声は自発性が高く，朗読音声と異なり音声の音響的特徴が話者や発話様式・発話意図などの影響を受け多様に変化するため，目標話者の限られた音声データのみで自然

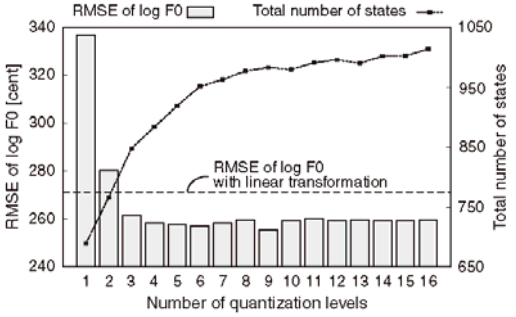

図 2 量子化レベル数と F0 歪及びモデル躁状態数の関係〔雑誌論文⑧〕

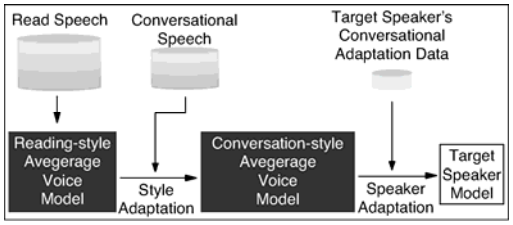

図 3 平均声モデルに基づく二段階モデル適応

性の高い合成音声を生成することは容易ではない。そこで，あらかじめ複数の話者の音声データを用いて学習された読上げスタイルの平均声モデルに対して，二段階のモデル適応を行う（図 3 参照）ことにより，目標話者の音声データがごく限られている場合でも目標話者の音響モデルを学習することが可能な手法を提案し，その有効性を示した〔雑誌論文⑦〕。

一方，上記手法によって合成音声の自然性が向上したとはいえ，実際に人間が発声した音声との差異は依然として大きいことも明らかになった。その要因の一つに，自発会話音声の韻律の多様性が十分に実現されていないことが挙げられる。これは HMM 音声合成において，読上げ調の音声を念頭に置いたコンテキストのみでは対話音声の多様性を表現するには十分でないことを意味している。

そこで本研究では，日本語話し言葉音声コーパス（CSJ）に含まれる様々な韻律情報をコンテキストとして新たに追加し，表 1 に示す拡張コンテキストセットを提案した〔学会発表③〕。さらに，コンテキストの増加による過学習を避けるための共有決定木クラスタリングにおける新たな停止基準の導入や，実用上のシステムを考慮して合成時に一部の追加コンテキストを自動推定する手法の提案を行い，その有効性を示した〔雑誌論文①〕。この他にも，新たな F0 モデリング手法の検討を行った〔学会発表①〕。

自発音声・会話音声合成の研究は，世界的にみても十分研究が進んでいるとは言えない状況であり，本研究で得られた知見は今後の研究の進展に役立つものと期待できる。

(3) パラ言語情報の検出・表出度合推定

研究代表者らは先に合成音声における感情表現や発話様式の表出度合を低次元のパラメータにより直観的に制御可能なスタイル制御と呼ぶ手法を確立した。本研究では，スタイル制御の逆過程を定式化し，入力音声に現れる感情表現や発話様式などのパラ言語情報の検出と表出度合を推定する手法を提案した〔雑誌論文⑩〕。提案手法は，音響特徴量を重回帰 HMM によりモデル化しておき，入力音声に対して重回帰モデルの説明変数の値を最尤推定することにより対象とするパラ言語情報を得ている。評価実験により，悲しみと喜びの二つの感情表現の推定では，線形回帰やサポートベクトルマシンと同等か上回る性能が得られ，読上げスタイルと学会講演スタイルといった発話様式の識別にも有効であることを示した〔雑誌論文⑩〕。

一方，従来の重回帰 HMM のモデル学習法では，二種類以上のスタイル音声が必要であること，重回帰モデルの学習データへの依存性が高いことといった問題点があった。これに対し，一種類のスタイル音声のみで学習可能かつ学習データのスタイルの偏りに対し頑健なモデル学習法を提案し，主観評価と重回帰説明変数の相関性を高める効果があることを示した〔学会発表④〕。

表 1 拡張コンテキストセット

BASELINE	ADDITIONAL
A 音素	F 音素引き延ばし
B モーラ	G 発話スタイル
C アクセント	H トーンラベル
D 呼気段落	I 非流暢性
E 発話長	J 音素付加情報
	K 単語単位
	L 節

(4) 話者・スタイル変動に頑健な音声認識

現在実用化されている音声認識システムは，ニュースなどの原稿読上げ調音声に対して人間に近い認識性能が得られるようになったものの，対話や講演音声などの自発性の高い自然発話音声に対してはまだ十分な性能であるとは言えない。これは音声の音響的特徴が言語情報だけでなく話者の発話様式や感情といったパラ言語情報によって大きく変動することが要因の一つとなっている。

これに対し本研究では，上述のスタイル推定手法に基づいた自然発話音声の発話様式識別と音声認識手法について検討を行った〔雑誌論文⑨，学会発表⑨〕。重回帰 HMM では，モデルの各分布の平均パラメータを低次元の説明変数ベクトル（スタイルベクトル）の重回帰により表現している。スタイルベクトルの各次元はそれぞれある一つのスタイルの表出度合に相当しており，入力音声に対してこれを最尤推定することによりスタイルの識別とその表出度合を推定することができる。また，スタイルベクトルを入力発話毎に推定し，推定されたスタイルベクトルを用いてモデルパラメータを変化させることにより，オンラインで音響モデルを入力音声のスタイルに適応することができる。

提案手法を CSJ の読上げスタイルと学会講演スタイルの音素認識に適用した結果，提案モデル適応手法は，スタイル依存やスタイル独立モデルに比べて高い音素認識率が得られることを明らかにした（図 4 参照）。また，一般話者の感情音声に対しても，話者非依存

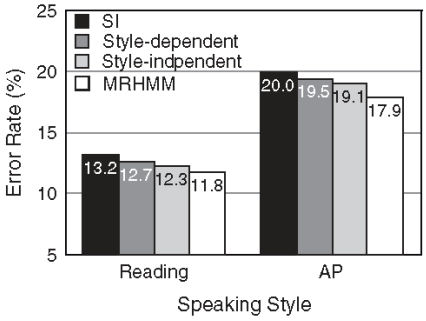

図 4 CSJ 読上げ(Reading)と学会講演(AP)音声に対する音素認識結果〔学会発表⑨〕

平静モデルから話者・スタイル適応を用いて重回帰 HMM を構成し，これにより高速にモデル適応が可能なことを示した〔雑誌論文⑨〕。

音声認識における話者適応に関する研究は数多いが，自発音声や感情音声のスタイル識別やモデル適応の研究は十分ではなく，本研究の成果は重要な意味を持つと思われる。

(5) 動作からのパラ言語情報の抽出

会話シーンにおいて，しぐさは重要なパラ言語情報となる。例えば，肯定や断定を表す「頷き」，否定や疑問を表す「首の傾げ」，思考中であることを表す「凝視」，さらにただ単に頷くだけでなく「強い頷き」，「ゆっくりした頷き」，「連続した頷き」といったその度合や動きの角度や時間，回数などが関係してくる。

本研究では，このような顔の動作に着目し，会話シーンを撮影した動画像から顔の姿勢を推定する方法を提案した〔学会発表⑦〕。提案手法では，動画像中の顔の特徴点の 2 次元座標と，3 次元フェイスモデル上の 3 次元座標間の透視 n 点問題を解くことにより，各軸方向における回転量と並進量の 6 次元パラメータを 1 フレームの特徴量として用いている。

提案手法は，実用化の観点からはまだ処理量や精度に問題が残っており，その解決に向けての検討や動作と音声からのパラ言語情報の統合手法についても今後の課題である。

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 50 件）

① 郡山知樹，能勢 隆，小林隆夫，HMM に基づく対話音声合成における多様な韻律生成のためのコンテクストの拡張，電子情報通信学会論文誌，査読有，Vol.J95-D, pp.597-607, 2012.

② Takashi Nose, Takao Kobayashi, Very low bit-rate F0 coding for phonetic vocoders using MSD-HMM with quantized F0 symbols, 査読有, Speech Communication, Vol.54, pp.384-392, 2012.

③ Vataya Chunwijitra, Takashi Nose, Takao Kobayashi, A tone-modeling technique using a quantized F0 context to improve tone correctness in average-voice-based speech synthesis, 査読有, Speech Communication, vol.54, pp.245-255, 2012.

④ Takashi Nose, Takao Kobayashi, Speaker-independent HMM-based voice conversion using adaptive quantization of the fundamental frequency, 査読有, Speech Communication, vol.53, pp.973-985, 2011.

⑤ Takashi Nose, Yuhei Ota, Takao Kobayashi, HMM-based voice conversion using quantized F0 context, IEICE Trans. on Information and Systems, 査読有, Vol.E93-D, pp.2483-2490, 2010.

⑥ Shuji Yokomizo, Takashi Nose, Takao Kobayashi, Evaluation of prosodic contextual factors for HMM-based speech synthesis, 査読有, Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, pp.430-433, 2010.

⑦ Tomoki Koriyama, Takashi Nose, Takao Kobayashi, Conversational spontaneous speech synthesis using average voice model, 査読有, Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, pp.853-856, 2010.

⑧ Takashi Nose, Koujirou Ooki, Takao Kobayashi, HMM-based speech synthesis with unsupervised labeling of accentual context based on F0 quantization and average voice model, 査読有, Proc. 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, pp.4622-4625, 2010.

⑨ Yusuke Ijima, Takashi Nose, Makoto Tachibana, Takao Kobayashi, A rapid model adaptation technique for emotional speech recognition with style estimation based on multiple-regression HMM, 査読有, IEICE Trans. on Information and Systems, Vol.E93-D, pp.107-115, 2010.

⑩ Takashi Nose, Takao Kobayashi, A technique for estimating intensity of emotional expressions and speaking styles in speech based on multiple-regression HSMM, 査読有, IEICE Trans. on Information and Systems, Vol.E93-D, pp.116-124, 2010.

〔学会発表〕（計 43 件）

① Tomoki Koriyama, An F0 modeling technique based on prosodic events for spontaneous speech synthesis, 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2012, 2012 年 3 月 29 日, Kyoto, Japan.
② 金川裕紀, HMM 音声合成における不特定話者スタイル変換の検討, 電子情報通信学会音声研究会, 2011 年 12 月 20 日, 芝浦工業大学, 東京都江東区.
③ Tomoki Koriyama, On the use of extended context for HMM-based spontaneous conversational speech synthesis, 12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011, 2011 年 8 月 30 日, Florence, Italy.
④ Takashi Nose, A perceptual expressivity modeling technique for speech synthesis based on multiple-regression HSMM, 12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011, 2011 年 8 月 28 日, Florence, Italy.
⑤ Takashi Nose, Very low bit-rate F0 coding for phonetic vocoder using MSD-HMM with quantized F0 context, 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, 2011 年 5 月 26 日, Prague, Czech Republic.
⑥ Vataya Chunwijitra, Tonal context labeling using quantized F0 symbols for improving tone correctness in average-voice-based speech synthesis, 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, 2011 年 5 月 24 日, Prague, Czech Republic.
⑦ 宮崎悠樹, 動画像からの顔の姿勢推定による非言語情報の取得, 画像電子学会第 250 回研究会, 2010 年 3 月 23 日, 崇城大学, 熊本市.
⑧ Takashi Nose, HMM-based speaker characteristics emphasis using average voice model, 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, 2009 年 9 月 10 日, Brighton, UK.
⑨ Yusuke Ijima, Speaking style adaptation for spontaneous speech recognition using multiple-regression HMM, 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, 2009 年 9 月 7 日, Brighton, UK.

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://www.kbys.ip.titech.ac.jp/

６．研究組織
(1)研究代表者
小林隆夫（KOBAYASHI TARO）
東京工業大学・大学院総合理工学研究科・教授
研究者番号：70153616

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
長橋　宏（NAGAHASHI HIROSHI）
東京工業大学・像情報工学研究所・教授
研究者番号：20143084

(4)研究協力者
能勢 隆（NOSE TAKASHI）
東京工業大学・大学院総合理工学研究科・助教
研究者番号：90550591