人にやさしい音声インタフェース

鹿野 清宏 *1
猿渡 洋 *1
河原 英紀 *4
西浦 敬信 *6
河原 達也 *2
武田 一哉 *3
徳田 恵一 *5
李 晃伸 *5

*1 奈良先端科学技術大学院大学
*2 京都大学
*3 名古屋大学
*4 和歌山大学
*5 名古屋工業大学
*6 立命館大学

プロジェクトの概要

音声は，人間どうしの意思疎通において日常的に利用されているメディアである．そこで，機械とのインタフェースにおいても利用できるようにするのは自然な発想であり，そのための研究は古くからなされてきた[2]．特に最近の携帯電話・カーナビを含む情報機器や家電製品は，機能が複雑になり，使いこなすのが容易でなくなっているので，より自然な対話的インタフェース（NUI：Natural User Interface）で利用できる技術が強く求められている．また，家庭内ロボットにそのようなインタフェースを備える試みもあるが，ここでも音声対話の能力は不可欠である．

音声認識・合成の技術は，1990 年代以降飛躍的に進歩し，実用化も進んでいるが，利用されている場面はまだ限定的である．その主な理由として，さまざまな利用者・利用環境に対する頑健性に乏しく，利用者の話し方に関する負担が大きいこと，およびさまざまなアプリケーションを構築するのに開発者のコストがかかることが挙げられる．したがって，高精度で頑健な音声認識・合成の基盤技術を研究開発し，汎用性が高く廉価な（できるだけフリーの）ソフトウェアとして提供することが求められている．

このような背景のもと，2003 年度から 5 年間，文部科学省のリーディングプロジェクト「e-Society 基盤ソフトウェア」の一環として，「ユーザ負担のない話者・環境適応性を実現する自然な音声対話処理技術」に関する研究開発が行われた．奈良先端科学技術大学院大学，京都大学，名古屋大学，和歌山大学，名古屋工業大学，立命館大学を主体とし，日立製作所，旭化成，松下電器産業，オムロン，松下電工の協力を得て実施された．大学が主導権を持った産学連携の形態といえる．

本プロジェクトでは，人にやさしい音声インタフェースを実現するために必要な，(i) 大語彙連続音声認識，(ii) 話者・環境適応，(iii) ハンズフリー音声認識，(iv) 多様な音声合成，(v) 音声対話プラットフォーム，などの基盤ソフトウェアの開発を行った．大語彙連続音声認識プログラム Julius [3] は音声認識の基盤であり，フリーソフトウェアとしての開発を継続し，対話システム向けの高精度化を行った．また，さまざまな機器で利用できるように，マイコンへの移植も行った．話者・環境適応とハンズフリー音声認識は，ユーザに事前の発声登録などの負担をかけない，マイクロフォンを意識させない上で重要な技術であり，公衆情報案内端末やロボットを想定して研究開発を行った．音声合成では，多様な声質の実現が重要であり，高精度音声分析合成系 STRAIGHT[6] を用いた声質変換プログラムと，HMM（隠れマルコフモデル）に基づく柔軟性の高い音声合成プログラム HTS [8] の発展を進めた．

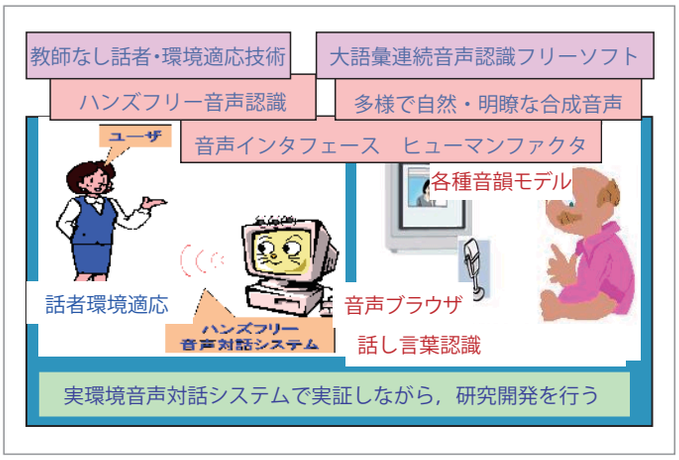

図 -1　人にやさしい自然な音声インタフェース

これらのプログラムを単に作成するだけでなく，さまざまなアプリケーションの音声対話システムを構築し，実環境で実証試験を長期間にわたって実施し，評価と継続的改善を行うとともに，応用に関する知見を蓄積した．

これに加えて，静かな音声メディアとして，音声を明示的に発声しない非可聴つぶやき [4] の認識・変換合成，さらに実環境での自然な入力のために，ハンズフリーのブラインド音源分離 [5] に関する研究も行った．

本プロジェクトの研究開発の概要を**図 -1** に示す．さらに，研究開発項目ごとの成果ソフトウェアを**表 -1** にまとめる．

研究開発項目	主な成果ソフトウェア
大語彙連続音声認識ソフトウェア	大語彙連続音声認識プログラム Julius 4.0 マイコン SH-4A 版 Julius 話し言葉に対応した言語モデル構築ツール
ユーザ負担のない話者・環境適応	話者に負担をかけないオンライン話者適応 非可聴つぶやきの認識・変換合成
ハンズフリー音声認識	ハンズフリー音声収録プログラムと DSP BSS 音源分離オンラインプログラム
多様な声質の音声合成ソフトウェア	高精度音声変換プログラム (STRAIGHT) HMM ベース音声合成システム (HTS)
実環境音声対話システム	たけまるくん音声対話システム構築キット 自動車内音声認識プログラム

表 -1　研究開発の成果

大語彙連続音声認識プログラム Julius

筆者らは，大語彙連続音声認識の汎用的なソフトウェア基盤として，オープンソースの音声認識エンジン Julius（http://julius.sourceforge.jp/）と標準的な音韻モデル・言語モデルを開発してきた．ただし，従来の「日本語ディクテーション基本ソフトウェア」[1] は，書き言葉調の文章の丁寧な読み上げ音声を主な対象としていた．これに対して，音声対話システムを指向して，さまざまな研究開発を行った．また，組込み機器にも利用できるように，日立製作所の協力のもと，マイコン（SH-4A）への実装も行った．**図 -2** に研究開発の概要を示す．

【 Julius 4.0 】

音声認識エンジン Julius について，以下に挙げるような対話システム向けの機能強化を行った．

- 非音声区間の棄却
- 音声区間の頑健な検出
- 複数の音響・言語モデルの併用・切替え
- 単語グラフの出力
- 認識結果の信頼度の付与

さらに，大幅な高速化や省メモリ化も実現し，マイコン（SH4-A ；400MHz,128MB）でも 2 万語彙の連続音声認識の実時間動作を可能にした．これらの集大成として，2007 年 12 月に約 8 年ぶりのメジャーバージョンアップとなる Julius 4.0 をリリースした．

【 Web テキストの選択的収集による言語モデルの半自動構築 】

音声対話による情報検索システムを構築するには，レストラン検索や観光案内などの当該アプリケーションのドメインに特化した言語モデルが不可欠であるが，人手で記述した有限状態文法は受理できる発話パターンが限定され，ユーザの負担が大きい．一方，頑健性・柔軟性の点で優れている統計的言語モデル（単語 N-gram）を構築するには，当該ドメインの学習用テキストを大量に用意する必要がある．そこで，Web テキストを収集し，当該ドメインに合致した話し言葉調の文を選択することで，アプリケーションに応じた言語モデルを効率的に構築する方法を研究し，ツールキットとして作成した．

教師なし話者・環境適応

誰もが音声認識を利用できるようにするためには，話者に負担をかけないユーザおよび発話環境への適応技術が不可欠である．そのために，事前のエンロール（発声登録）を一切必要としない，教師なし話者・環境適応に関する研究開発を行った．さらに，静かな音声メディア非可聴つぶやきを新たに見出し，声を出さない音声認識（無音声認識）や声を出さない音声通話（無音声電話）の研究も行った．

【 教師なし話者適応と音韻モデルの構築 】

公衆の場に設置された音声対話システムを想定して，1 文程度の発声で音韻モデルを教師なしで適応する方式の研究開発を進めた．HMM の十分統計量を効率的に活用する話者適応アルゴリズムを考案し，約 5 秒で話者適応を実現した．

また，既存の音声データベースを活用できる十分統計量に基づく音声データ選択アルゴリズムを考案して，新たな環境における音韻モデルを効率よく構築する方式を実現した．

【 非可聴つぶやき（NAM）】

公衆の場で音声認識を利用する際には，背景雑音の問題もあるが，周囲を配慮して大声で発声するのが憚られ

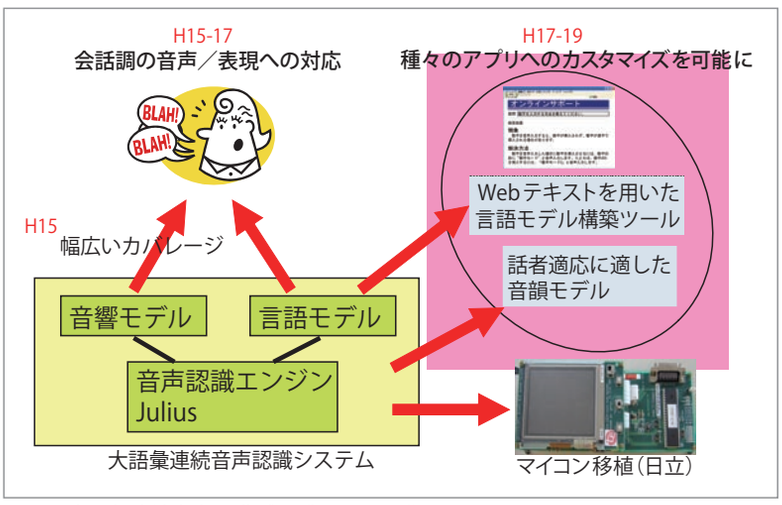

図 -2　大語彙連続音声認識プログラムの開発

ることも考えられる．この問題に対して，静かな音声メディアとして非可聴つぶやき（NAM：Non-Audible Murmur）を見出した[4]．これは，声を出さない状態でつぶやくように発話して，耳の後ろに置いたセンサで波形を取り込むものである．これを音声認識と同様の枠組みで認識したり，通常の音声に変換・合成することにより無音声電話を実現する可能性を示した．またこれに基づいて，発話障害者補助の研究も立ち上げた．

ハンズフリー音声認識

ユーザに負担をかけない自然な音声入力系として，ハンズフリー音声認識システムを構築した．特に，比較的コンパクトかつ廉価なマイクロフォンアレイシステムを用いて，音声認識性能の向上を目指した．具体的には，ユーザからの距離 1m 以下で高性能に動作するハンズフリー音声認識システムを，8 チャネル以下のマイクロフォンアレイを用いて開発することを目標とした．さらに，マイクロフォンアレイのコストを下げるため，ハンズフリー音声収録用 DSP も開発した．

【 SSA：空間スペクトル演算アレイ 】

本研究開発では，効率的かつ高精度な雑音抑圧アルゴリズムとして空間スペクトル演算アレイ SSA（Spatial Subtraction Array）を提案・実装した．一般に実環境で収録される音は，認識対象である目的音声成分と不要な雑音成分の和で表される．SSA は，主に目的音声を強調する「主パス」，雑音のみを推定する「参照パス」から構成される．遅延和アレイ信号処理による主パスでは，目的音声を強調するが，雑音成分が残留する．その残留雑音成分の近似値を参照パスで推定する．ここでの近似値とは，複素信号における振幅値のみ等しく，位相値は無視したものである．両パスの信号をパワースペクトルドメインで減算（各周波数帯域で位相成分を無視して二乗振幅成分間で減算）することにより，従来の線形信号処理以上の雑音抑圧効果を得ることができる．また，音声認識で必要とされない位相成分の復元を行わないため，高速な処理が実現できるという特徴がある．

本 SSA 処理の前処理として，音声の特徴量および空間的情報を考慮した重み付き CSP 法に基づく実時間方位推定･発話検出法を導入し，音声検出精度の向上を図った．さらに，DSP モジュール上に実時間 SSA 処理系を実装し，ハンズフリー音声対話デモシステムをコンパクトなハードウェア上に構築した（**図 -3** 参照）．

【 BSS：ブラインド音源分離 】

本システム開発と並行して，音源間の独立性のみに基づいて分離を行う歪みなしのブラインド音源分離（BSS）の研究[5]を行った．さらに，この BSS がマイクロフォンの誤差や室内残響に影響されることなく雑音推定を行えることに着目し，これを上記 SSA の雑音推定部（参照パス）に導入した BSSA（ブラインド SSA）を開発した．最終的に，駅に設置された音声対話システム（後述）に BSSA をリアルタイム実装し，ハンズフリー環境における高精度な音声認識システムの開発に成功した．

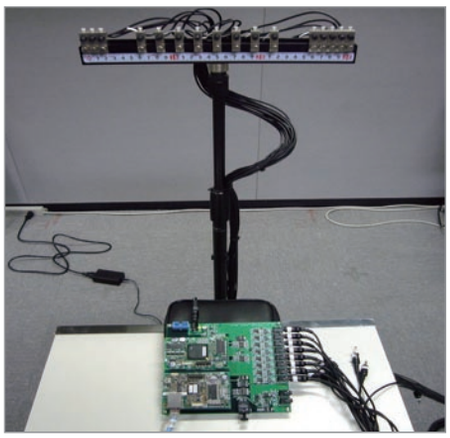

図 -3　マイクロフォンアレイおよび SSA 用実時間 DSP

多様な音声合成（STRAIGHT と HTS）

機械と人間との対話を自然なものにするためには，感情や話し方に応じて，多様な声質の音声を自由に合成できる必要がある．そのためには，音声パラメータの精密な分析・変換・再合成の要素技術と，それらのパラメータの統計的性質を利用した音声合成技術が必要となる．本プロジェクトでは，前者の要素技術として STRAIGHT を，後者として HMM を用いた音声合成システム（HTS）を採用し，それらの開発に必要なデータベースの構築と併せて，両者の発展と統合を進めた．

本プロジェクトは，それまで独立に進められてきた音声分析変換合成の STRAIGHT[6]と，統計的パラメトリック音声合成の HTS[8]に連携・協力の場を提供し，それらの研究を大きく促進することに貢献した．その結果，世界の最先端レベルの音声合成技術が確立されるとともに，研究機関・企業にとって利用しやすい，応用システム開発のための基盤ソフトウェアが整備された．

【 STRAIGHT：高精度音声モーフィング 】

要素技術としての STRAIGHT では，利用形態として，リアルタイムシステム[7]と高品質なオフライン作業用のシステムを想定した．それ自身が研究対象であったことから採用されていた Matlab での実装を見直し，応用システム開発に利用しやすい C 言語によ

りプログラムを実装した．リアルタイム版では，STRAIGHT のプログラムの制御構造を根本から見直し，実時間向きのアルゴリズムの採用や効率化により，標準的な PC での実時間動作を実現した．オフライン版では，応用システム開発を容易にすることを目的として，STRAIGHT のプログラムインタフェースを合理的に設計し直した．また，構築したデータベースを用いることにより，チューニングと高品質化のための改良を進めた．これらを STRAIGHT suite（http://straight-suite.sys.wakayama-u.ac.jp/）として公開した．

図 -4　駅に設置した音声情報案内システム「キタちゃん」（左）とキタロボ（右）

【 HTS：HMM ベース音声合成システム 】

音声合成システム HTS では，開発のマイルストーンとして，HMM を用いた音声合成システムに，STRAIGHT およびいくつかの改良手法を導入することにより，音声合成国際評価会 Blizzard Challenge のためのシステムを構築した．本システムは，これまでの常識を覆す合成音声の品質向上に成功しており，統計的パラメトリック音声合成方式に対する考え方を一変させるきっかけとなった．これらの成果の社会還元のため，本システムの関連ソフトウェアを HTS（http://hts.sp.nitech.ac.jp/）として公開した．この HTS と STRAIGHT を組み合わせることにより，世界最先端レベルの音声合成システムの構築が可能になった．

こうした研究開発は，すでにいくつもの波及効果を及ぼしている．その典型例として，非可聴つぶやきによる音声コミュニケーションが，音声変換合成技術と組み合わせることで生み出された [9]．

音声対話システムのフィールドテスト

利用者が感じる音声認識システムの性能は，SN 比や発話速度といった計測可能な音声の性質だけでは説明が難しい多様な要因によって支配されている．実際の応用場面において，多様な利用者・利用形態における音声認識システムの振舞いを記録し，アルゴリズムの評価やエラーの解析を行うことは，音声対話システムの高度化に欠かすことのできないステップである．本プロジェクトでは，以下のように，多様な技術を実装したシステムを用いてフィールドテストを行った．

(1) **公共施設案内システム**：本プロジェクト開始時から，生駒市北コミュニティセンターに音声情報案内システム「たけまるくん」を設置し，5 年間にわたって運用を行った．その後，学研北生駒駅に「キタロボ」「キタちゃん」（**図 -4**）を設置し，2 年間運用した．この間，幼児から高齢者までの多様な利用者の音声を収録することができた [10]．この音声データベースを整備して，音韻モデルや言語モデルの改善に用いた．なお，本データベースは公開を開始している．

(2) **バス運行情報案内システム**：京都市バスのリアルタイムの運行情報を電話音声で案内するシステムを構築し，本プロジェクト開始時から約 4 年間にわたって運用した．長期間にわたる一般ユーザの対話システムの利用に関する知見が得られた．

(3) **観光情報案内システム**：ドメインを限定した大規模文書（Wikipedia の京都関連エントリ）に基づいて，情報検索・質問応答・情報推薦を行う会話ロボットを構築し，3 カ月間にわたって大学の博物館で運用した [11]．

(4) **楽曲検索システム**：PC 内にある音楽を検索・再生するシステムの音声対話インタフェースを配布し，家庭やオフィスなどのさまざまな PC 操作環境下で利用実験を行った [12]．利用者が持つ楽曲にあわせて，自動的に認識辞書を構築しダウンロードできる機能など，インターネット環境を想定した遠隔制御技術の実験を行った．さらに，走行自動車内で楽曲検索を行うシステムにも発展させた．

これらのフィールドテストの結果，実環境下で発声された大量の音声データが収集され，子供の音声の音響的特徴，幼児の発声変形，対話文の言い回し等に関する統計的なモデルの高精度化が可能となった．頑健な発話検出方法，Web テキストと対話文を併用する言語モデルの構築方法，発話語彙に合わせた音響モデルの構築方法といった新しい技術が開発され，それらの有効性が確認された．一方で，利用者が主観的に感じる音声認識の性能と，実際の音声認識率とは必ずしも高い相関がない（相関係数 0.4）など，利用者の利便性向上にはさらなる研究が必要であることも明らかになった．

まとめ

本研究開発による成果を以下にまとめる．

(1)「たけまるくん」などの多くの音声情報案内システムを実環境で運用して，それらの有効性を示すとともに，幼児から高齢者までの音声データの収集や話者適応技術開発により，音声認識の性能の向上を達成した．

(2) 歪みなしのブラインド音源分離と実環境音声データによる音韻モデルにより，1m 以上離れた位置からのハンズフリー音声認識の性能を，接話マイクと同等のレベルまで高めた．

(3) 大語彙連続音声認識プログラム Julius に，雑音に頑健な手法などを採り入れて，Julius 4.0 として配布を開始した．また，音響モデルや言語モデルの話し言葉対応も進めて，大学の講義や議会（国会）の音声認識の性能を飛躍的に向上させ，音声認識の適用分野を広げた．さらに，Julius のマイコン SH-4A への実装も進め，2 万語彙の実時間での連続音声認識を実現した．

(4) 音声分析合成プログラム STRAIGHT の改良を行い，精度の向上とともに大幅な計算量の削減を達成し，実時間の高品質音声モーフィングが可能になった．HMM を用いた音声合成システム HTS も改良を重ねて，国際評価会を通して性能の高さが世界に知られ，多くの研究機関や企業で使われるようになった．

(5) 非可聴つぶやき（NAM）や歪みなしのブラインド音源分離（BSS）により，騒がしい場所でのハンズフリー通話や音声認識が可能になった．NAM による発話障害者の発話補助や BSS による両耳補聴器の可能性が示され，音声のユニバーサルコミュニケーションの研究領域を大幅に広げることができた．

謝辞　本稿で述べた研究開発は，文部科学省のリーディングプロジェクト「ユーザ負担のない話者・環境適応性を実現する自然な音声対話処理技術」で実施されたものである．本プロジェクトに参加・協力いただいた奈良先端科学技術大学院大学，京都大学，名古屋大学，和歌山大学，名古屋工業大学，立命館大学，日立製作所，旭化成，松下電器産業，オムロン，松下電工の皆様に感謝します．

参考文献

1）鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄：音声認識システム，オーム社(2001)．
2）河原達也，荒木雅弘：音声対話システム，オーム社(2006)．
3）河原達也，李　晃伸：連続音声認識ソフトウェア Julius，人工知能学会誌，Vol.20, No.1, pp.41-49 (2005).
4）中島淑貴，柏岡秀紀，ニックキャンベル，鹿野清宏：非可聴つぶやき認識，電子情報通信学会論文誌，Vol.J87-D-II, No.9, pp.1757-1764 (2004).
5）Takatani, T., Nishikawa, T., Saruwatari, H. and Shikano, K. : High-fidelity Blind Separation of Acoustic Signals Using SIMO-model-based Independent Component Analysis, IEICE Transactions on Fundamentals, Vol.E87-A, No.8, pp.2063-2072 (2004).
6）河原英紀：Vocoder のもう 1 つの可能性を探る―音声分析変換合成システム STRAIGHT の背景と展開―，日本音響学会誌，Vol.63, No.8, pp.442-449 (2007).
7）坂野秀樹，森勢将雅，高橋　徹，西村竜一，入野俊夫，河原英紀：リアルタイム STRAIGHT の改良と STRAIGHT ライブラリの実装，電子情報通信学会技術研究報告，SP2007-213, pp.157-162 (2008).
8）Zen, H., Toda, T., Nakamura, M. and Tokuda, K. : Details of the Nitech HMM-based Speech Synthesis System for the Blizzard Challenge 2005, IEICE Trans, Information and Systems, Vol.E90-D, No.1, pp.325-333 (2007).
9）中村圭吾，戸田智基，猿渡　洋，鹿野清宏：肉伝導人工音声の変換に基づく喉頭全摘出者のための音声コミュニケーション支援システム，電子情報通信学会論文誌，Vol.J90-D, No.3, pp.780-787 (2007).
10）Cincarek, T., Kawanami, H., Nishimura, R., Lee, A., Saruwatari, H. and Shikano, K. : Development, Long-Term Operation and Portability of a Real-Environment Speech-oriented Guidance System, IEICE Trans, Information and Systems (2008).
11）翠　輝久，河原達也，正司哲朗，美濃導彦：質問応答・情報推薦機能を備えた音声による情報案内システム，情報処理学会論文誌，Vol.48, No.12, pp.3602-3611 (2007).
12）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，電子情報通信学会論文誌 D，Vol.J90-D, No.10, pp.2807-2816 (2007).

（平成 20 年 8 月 2 日受付）

鹿野　清宏(正会員)　shikano@is.naist.jp
奈良先端科学技術大学院大学教授．昭 45 年名大・工・電気卒．昭 47 年同大修士了．工博．電電公社武蔵野電気通信研究所，ATR 自動翻訳電話研究所などを経て現職．音声・音情報処理の研究および研究指導に従事．本会フェロー．

河原　達也(正会員)　kawahara@i.kyoto-u.ac.jp
京都大学学術情報メディアセンター教授．同大博士（工学）．音声認識・理解および音声対話システムに関する研究に従事．本会音声言語情報処理(SLP)研究会主査．

猿渡　　洋　sawatari@is.naist.jp
奈良先端科学技術大学院大学准教授．名大博士（工学）．音声・音響信号処理および音声対話システムに関する研究に従事．

武田　一哉(正会員)　kazuya.takeda@nagoya-u.jp
名古屋大学情報科学研究科教授．同大博士（工学）．音声信号処理・音声認識および音声対話システムに関する研究に従事．本会音声言語情報処理(SLP)研究会前(H18 ～ 19)主査．

河原　英紀(正会員)　kawahara@sys.wakayama-u.ac.jp
和歌山大学システム工学部教授．北大博士（工学）．音声分析・変換・合成および聴覚メディア処理に関する研究に従事．

徳田　恵一(正会員)　tokuda@nitech.ac.jp
名古屋工業大学大学院工学研究科教授．平元東工大大学院総合理工学研究科博士課程修了．音声言語情報処理，統計的学習理論他の研究に従事．

西浦　敬信(正会員)　nishiura@is.ritsumei.ac.jp
立命館大学情報理工学部准教授．博士（工学）奈良先端科学技術大学院大学．音響信号処理，主として音環境の解析，理解，再現，合成に関する研究に従事．

李　　晃伸(正会員)　ri@nitech.ac.jp
名古屋工業大学大学院工学研究科准教授．京大博士（情報学）．大語彙連続音声認識・音声言語理解・音声対話システム・音声インタフェースに関する研究に従事．