サービスや利用環境の変化に柔軟に対応する音声認識技術

音声認識技術は，人の声を機械が聞き取る技術です．音声認識は，音声言語インタフェースの入り口にあたる技術で，インタフェース全体の性能を左右するため，常に高い精度で認識することが重要ですが，現在の音声認識技術はサービスの種類や利用環境によっては認識精度が大きく低下する場合があります．本稿では，その課題を克服するための最新の取り組みを紹介します．

まさたき ひろかず†1 あさみ たいち†1
政瀧 浩和 /浅見 太一
やまはた しょうこ†1 ふじもと まさきよ†2
山畠 祥子 /藤本 雅清

NTTメディアインテリジェンス研究所 †1
NTTコミュニケーション科学基礎研究所 †2

音声認識によるユーザインタフェース

音声認識技術は，機械が人間の言葉を聞き取って文字にする技術，すなわち人間の耳に相当する能力を機械で実現する技術です．そして同時に私たちが提案する音声言語インタフェースの入り口を担う技術でもあり，常に高い精度で人間の声を聞き取ることが求められます．

音声認識技術はNTT研究所で40年以上の研究開発の歴史があります．音声認識手法の技術革新と計算機の性能向上により，2000年ごろには特定の条件ではすでに実用的な技術になっていました[(1)，(2)]．音声認識技術の標準的な構成を**図1**に示します．音声認識技術は，音声特徴量抽出，音響モデル，認識辞書の3つの部品から構成されています．現状の技術では，静かな場所で，標準的な声の持ち主が，一般的な単語をしゃべる場合は精度良く認識してくれますが，次のような場合

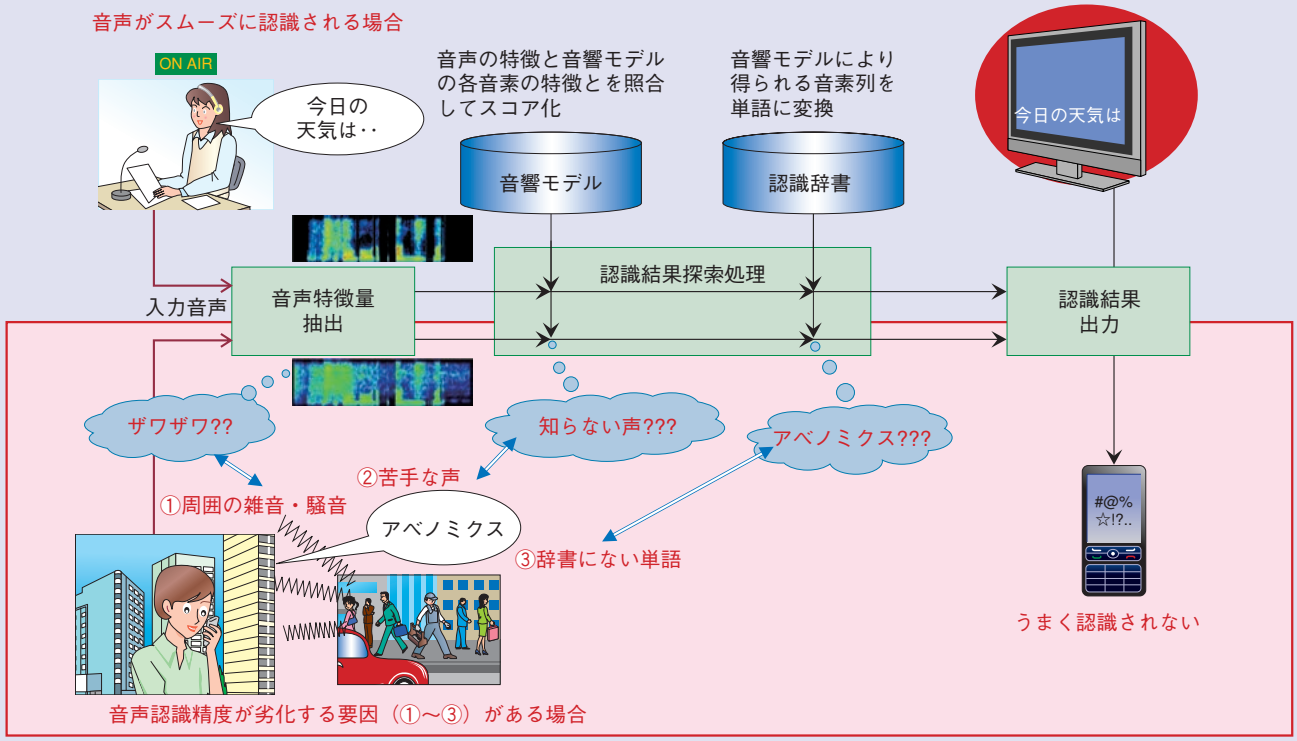

図1　音声認識技術の原理と課題

は認識精度が劣化します．

① 周囲の雑音・騒音：人ごみや駅の構内など，騒音の激しい場所でしゃべると，入力音声に人間の声と騒音とが混ざることにより音声の特徴が乱れ，認識精度が大きく下がります．

② 苦手な声の存在：音響モデルは，多数の人の声を学習して作成されますが，学習した声と特徴が異なる声はうまく認識できない場合があります．

③ 認識辞書にない単語の発声：現在の音声認識技術では，辞書にない単語は認識してくれません．最近生まれた新しい単語や，一般の人があまり使わない単語は，認識辞書に登録されていない場合があり，正しい認識結果を出力することはできません．

以上のように，現在の音声認識技術は使う環境，使う人の声質，話す内容等の変化に弱いのですが，我々はこれらの問題を克服することを目指して，日々研究開発に取り組んでいます．

音声区間抽出・雑音抑圧技術

周囲の雑音が大きい場所でしゃべった場合，なかなか正しく音声を認識してくれません．この問題を解決するためには，音声区間検出（Voice Identification）と，雑音抑圧（De-noising）という2つの技術を利用することが必要不可欠となります．音声区間検出技術は，雑音で埋もれた入力音データの中から人間が発話した時間区間を特定する技術であり，雑音抑圧技術は，特定された発話区間の音データから雑音を取り除き，人間の音声のみをクリアに取り出す技術です．

これまでは，音声区間検出と雑音抑圧を個別に用意し，各々の技術は単純に音データの入出力関係のみで連結されていました．しかしこの場合，各技術で発生する誤りが蓄積し，処理が進むにつれて性能が劣化するおそれがあります．

このような技術において我々は，音声区間検出と雑音抑圧を統合し，各々の処理を同時に実行可能とする画期的な技術，DIVIDE（Dynamic Integration of Voice Identification and DE-noising）を開発し，技術統合により各技術で発生する誤りを軽減し，かつ相補的に性能改善が得られることを示しました[(3)]．DIVIDEでは音声認識と同様に，人間の音声を学習した統計モデルを活用しており，**図2**に示すように，この統計モデルを音声区間検出と雑音抑圧で情報共有することにより，高度な処理を行っています．具体的には，統計モデルを利用して，ある一定の時間区間ごとの入力音データに人間の音声が含まれる確率を計算します．そして確率がしきい値以上であれば，その時間区間を人間の発話区間として特定します．またそのとき，雑音抑圧においても人間の音声が含まれる確率を利用して，人の発話が存在する時間区間と存在しない時間区間を区別しながら雑音を抑圧することにより，高い音声品質を得ることが可能となっています．すなわち，人間の音声の統計モデルという知識を活用して，ある時間区間における入力音データが人間の音声らしいか，らしくないかという情報を抽出し，その情報に基づいて音声区間検出と雑音抑圧を同時に実行しています．

図3は，雑音がある環境での音声認識の結果を示しており，DIVIDEにより音声認識率が大幅に改善されていることが確認できます．

苦手な声への自律適応技術

音声認識では，私たちの発した声と音素*とを対応付ける，「音響モデル」を利用します．声と音素との対応は，人が耳で聞いて声を文字に起こした

＊音素：私たちの声を構成する最小単位で，単語をローマ字で書いたときのアルファベットの単位にほぼ相当します．

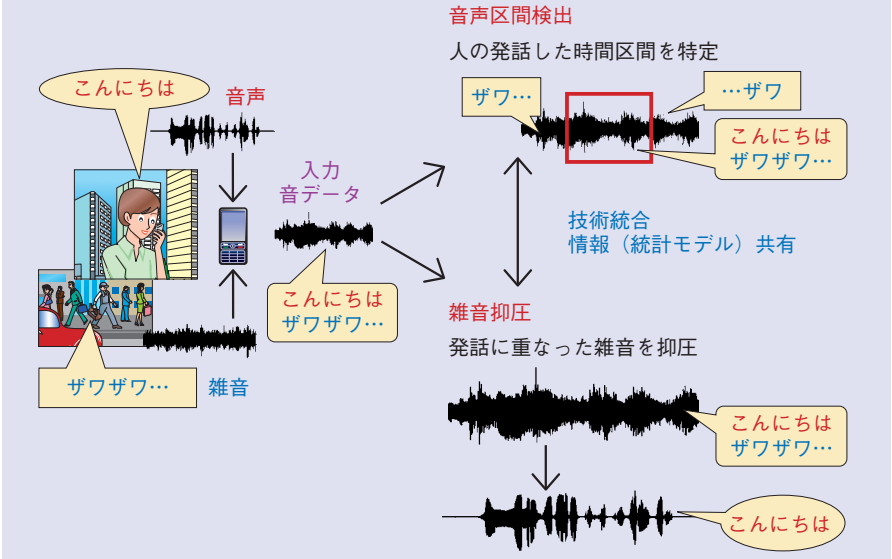

図2　DIVIDEの概要

「正解テキスト」に基づいて音響モデルに学習されます（**図 4** ①）．人の声は十人十色ですので，音響モデルにはあらかじめ多数の話者の声と音素の対応を学習させています．たとえ自分の声そのものが学習されていなくても，自分と特徴の似た声が学習されていれば，音響モデルは正しく音素と対応付けることができるため，多くのユーザの声を高精度で認識することが可能になっています．

しかし，多数の話者の声を学習させたとしても完全に死角をなくすことは困難です．実際の音声認識サービス利用シーンでは，音声認識精度が大きく落ちてしまう一部のユーザ（苦手話者）がどうしても存在します．どういった声で認識精度が落ちるかは，実際にその人がサービスを利用しはじめるまで分からないため，事前の学習によってできるだけ多くのユーザをカバーしようとする従来のアプローチで予防するには限界がありました．

この問題を解決し，より多くのユーザに対応できる音声認識を実現するため，苦手な声への自律適応技術を開発しました．これは，稼動中の音声認識エンジン自身が苦手話者の出現に気付

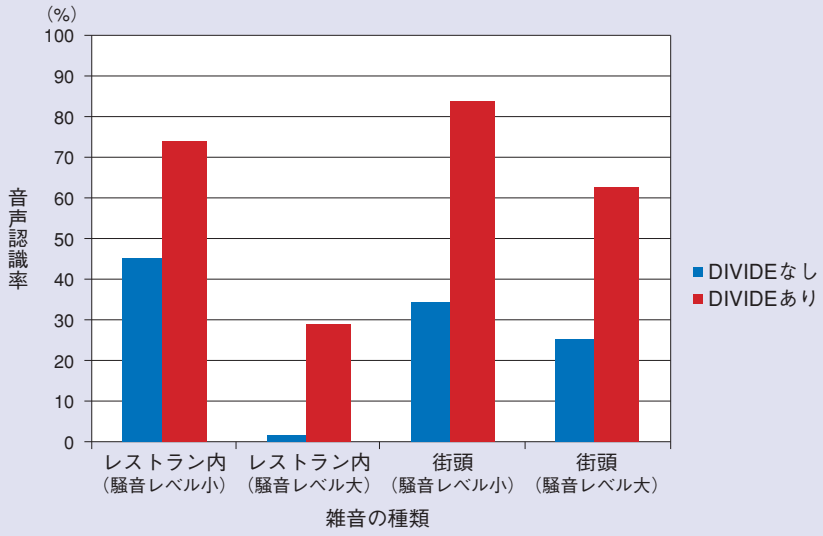

図 3　DIVIDE による音声認識結果

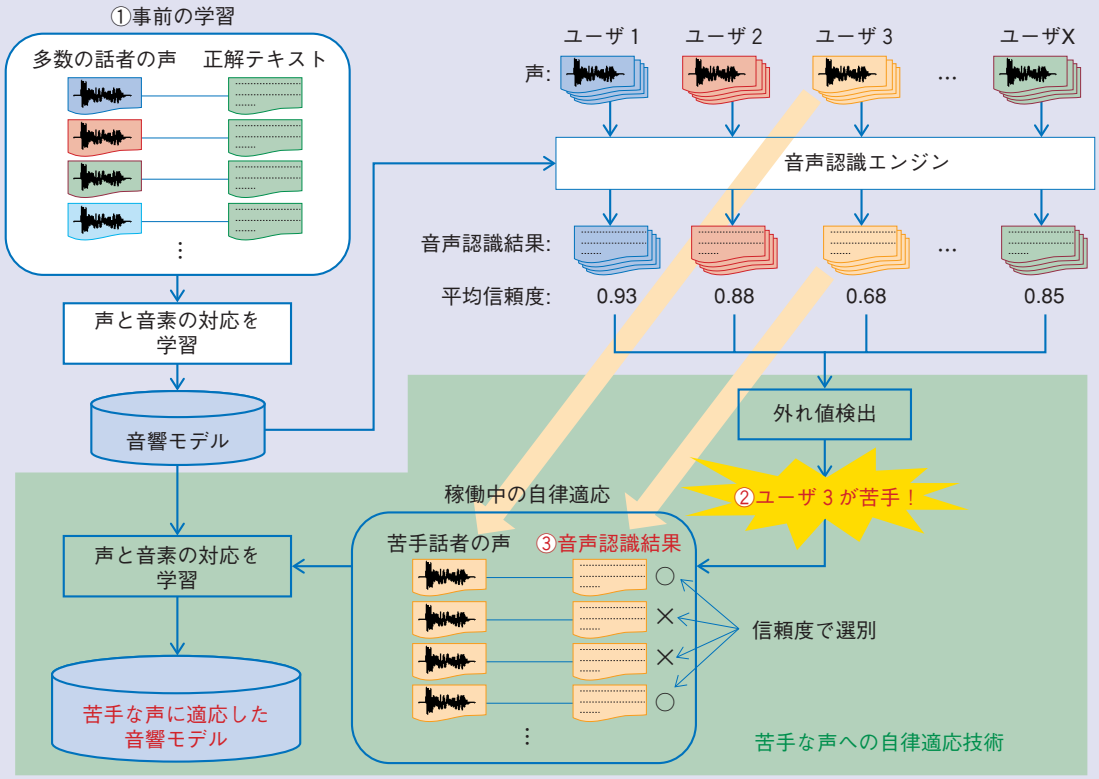

図 4　苦手な声への自律適応技術の概要

き，自動的にその人の声を音響モデルに取り入れることで，高い認識精度を維持する技術です（図４）．

音声認識エンジンは，自らの出力した音声認識結果がどの程度正しそうかを自己採点し，スコア付けする機能を持っています[4]．この自己採点スコア（信頼度）を利用して，苦手話者を自動的に検出することが可能です．音声認識を利用するいろいろなユーザの中で，ある特定のユーザにおいて平均的に信頼度が低くなり，外れ値となっている場合，そのユーザは苦手話者だと考えられます（図４②）．

検出した苦手話者の声と音素との対応を新たに音響モデルに学習させる際，通常はその話者の声に対応する正解テキストが必要となります．しかし，正解テキストの作成には非常に手間がかかるため，サービスの運用中に苦手話者を見つけるたびに正解テキストを作成することは実際には困難です．そこで，自動的に苦手話者の声を音響モデルに学習させるため，この技術では正解テキストの代わりに，音声認識エンジン自らが出力した音声認識結果を利用します（図４③）．正解テキストとは違い，音声認識結果には実際の発声内容とは異なる誤った部分が含まれているため，そのまま使うと誤った声と音素の対応を学習してしまい，良い効果が得られません．そこで，ここでも信頼度を活用し，しきい値処理によって誤りの少ない音声認識結果を選別したうえで学習に使う工夫をしています．

評価実験の結果，サービスを利用するユーザに合わせて音響モデルを変化させ，それまで苦手とされていた話者のうち80％は通常どおりの精度まで改善できることを確認しています．

新しい語彙への自律適応技術

音声認識システムは「認識辞書」という認識できる単語のリストを持っており，これに登録されていない単語（未知語）は認識することができません．しかし，会話の中で出てくる重要なキーワードは，アーティストの名前や，本のタイトルなど，すべての単語をあらかじめ辞書でカバーしておくには限界があります．このような単語を，ユーザがいちいち自分で追加しなければならないのはとても面倒です．

これに対し，従来とられてきた方法があります．それはWebページからサービスに関連しそうな文書を収集し，その中から未知語を抽出，認識辞書へ追加する，というものです（**図５**(a)）．しかし，文書に現れる未知語の中には，実際に話されるものと全く話されないものとが混在しており，すべての未知語を追加すると認識辞書にムダな単語が増えてしまいます．そして，これらのムダな単語は誤って認識されてしまい，正確な音声認識を阻害する要因になってしまいます．

そこで，収集した関連文書に出現する未知語の中から，実際に話されそうな単語のみを選別し，認識辞書に追加する技術を開発しました[5]（図５(b)）．これにより，ムダな単語が認識誤りを

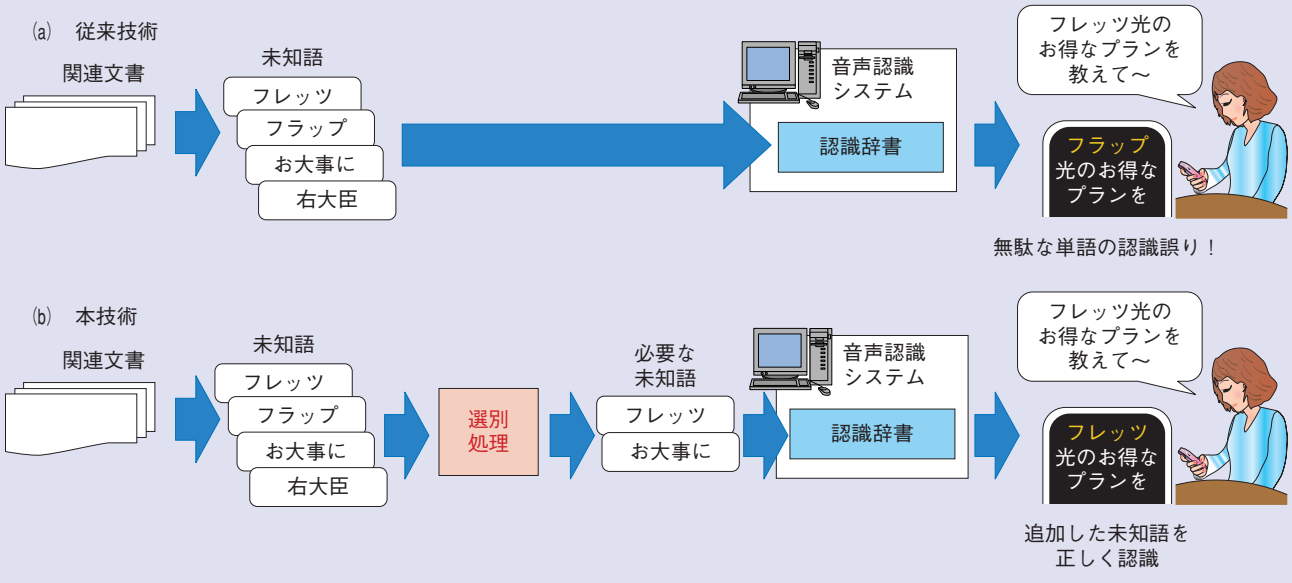

図５　従来技術と本技術の比較

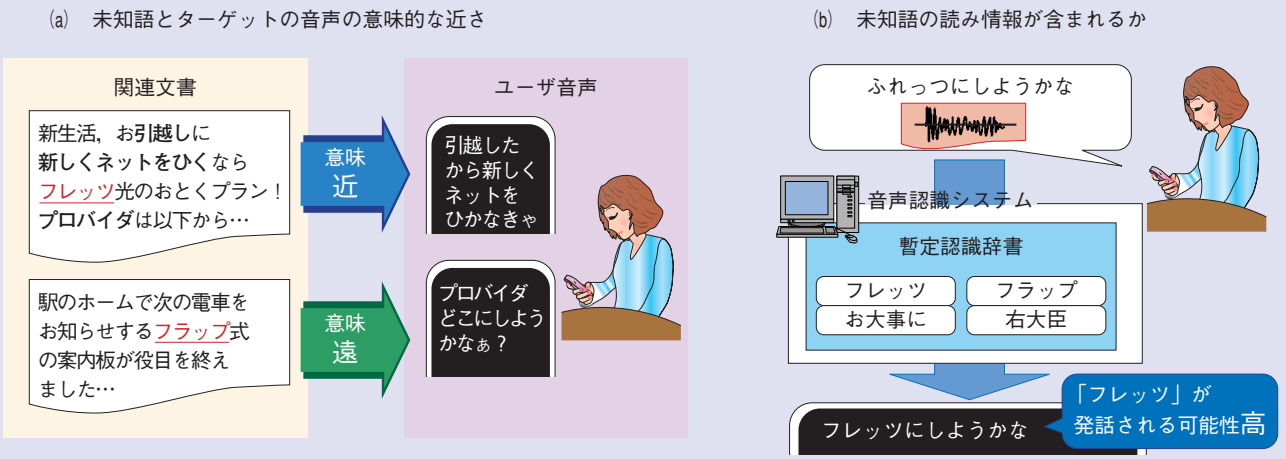

図6　本技術による未知語選択の仕組み

起こすことのない，ユーザやサービスに合った認識辞書をつくることができます．

この未知語の選別には2つの情報を用います．1つは，未知語とターゲットの音声の意味的な近さです（**図6** (a)）．例えば，ユーザの音声を認識したとき，「引越し」や「ネット」といったフレッツ光の利用シーンに関係の深い単語が多く出現したとします．すると，関連文書中の未知語で，「引越し」や「ネット」といった単語と一緒に出現（共起）する未知語を優先的に選出します．もう1つは，未知語の読み情報に近い音が，ターゲットの音声に含まれているかどうかです（図6 (b)）．例えば，「ふれっつひかりにしようかな」という音声があれば，「フレッツ」は音声の中で発話されている可能性が高いと考えられます．そこで，まずすべての未知語を追加した暫定認識辞書を作成し，音声認識を行います．そして，認識結果に含まれる未知語を優先的に選出します．この2つの情報を統合することにより，実際に発話される未知語だけを高い精度で選別することができます．

この技術により，従来のすべての単語を追加する方法よりも，無駄な単語が誤って認識される割合を約10％削減することができました．

今後の展開

我々は，今までNTT研究所で培ってきた音声認識技術をVoiceRex（ボイス・レックス）というエンジンとして開発してきました．今回紹介した技術も近い将来VoiceRexに搭載される予定です．今後は，環境・声の特徴・単語の特徴を個人レベルまでに特化，すなわちパーソナライズ化させることにより，認識精度をさらに高めることを検討しています．また，音声認識技術をクラウド上で動作させることにより，音声認識をより身近にし，「誰でも・いつでも・どこでも」使える技術にしたいと考えています．

■参考文献

(1) 野田・山口・大附・今村：“音声認識エンジンVoiceRexを開発，”NTT技術ジャーナル，Vol.11，No.12，pp.14-17，1999.
(2) 政瀧・柴田・中澤・小橋川・小川・大附：“顧客との自然な会話を聞き取る自由発話音声認識技術「VoiceRex」，”NTT技術ジャーナル，Vol.18，No.11，pp.15-18，2006.
(3) M. Fujimoto，K. Ishizuka，and T. Nakatani：“A study of mutual front-end processing method based on statistical model for noise robust speech recognition，”INTERSPEECH 2009，pp.1235-1238，Brighton，U.K.，Sept. 2009.
(4) T. Asami，N. Nomoto，S. Kobashikawa，Y. Yamaguchi，H. Masataki，and S. Takahashi：“Spoken document confidence estimation using contextual coherence，”INTERSPEECH 2011，pp.1961-1964，Florence，Itary，Aug. 2011.
(5) S. Yamahata，Y. Yamaguchi，A. Ogawa，H . Masataki，O. Yoshioka，and S. Takahashi：“Automatic Vocabulary Adaptation Based on Semantic Similarity and Speech Recognition Confidence Measure，”INTERSPEECH 2012，Portland，U.S.A.，Sept. 2012.

（左から）山畠　祥子／政瀧　浩和／浅見　太一／藤本　雅清

現状では利用範囲には制約がある音声認識技術ですが，今後さまざまな声が聞き取れるようになり，ますます皆様が身近に使ってもらえる技術となることを目指して，今後も研究開発に励みたいと思います．

◆問い合わせ先
NTTメディアインテリジェンス研究所
音声言語メディアプロジェクト
TEL　046-859-3004
FAX　046-855-1054
E-mail　masataki.hirokazu@lab.ntt.co.jp