7 音声インタフェースの現状とイノベーションの可能性

西村 雅史　　倉田 岳人
日本アイ・ビー・エム（株）東京基礎研究所

国内外における音声インタフェースの現状

音声認識，音声合成といった技術に基づく音声インタフェースは古くは1970年代後半頃から実用化されている．特に，物流の仕分けといった“アイズビジー・ハンズビジー”状況への対応（1978）や，ダイヤルパルス回線での電話音声自動応答サービスの実現（1982）など，先駆的な応用例の多くは日本で生み出された．その後は，試行錯誤が続いたが，主に障害者のアクセシビリティ改善などを目的として音声インタフェースの利用は徐々に広まっていった．

●音声インタフェースの成功例

80年代後半から90年代前半にかけてDARPA（米国防省高等研究計画局）が行ったディクテーション関連プロジェクト（大語彙連続音声認識による音声テキスト変換技術）の成果をベースとして，90年代後半になるとPC用のディクテーションソフトが世界的に普及した．結果的にはキーボードの代替として広く一般に用いられるという状況にはならなかったものの，診断報告書やカルテの作成など，医療分野を中心とした特定のアプリケーションでは欧米を中心に旺盛な需要があり，ビジネス的にも大きな成功を収めている．また，今後電子カルテの普及に伴ってさらに利用が進むとの予想もある☆1．

また，米国では90年代後半以降，こちらもDARPAプロジェクトによって生み出されたQ&A対話処理技術をベースとして，高度な電話音声自動応答サービスが実現され，多くのコールセンタが音声認識・音声合成技術を使って自動化された☆2, 1)．そして2000年以降には音声によるWebサーフィンを実現する技術として，音声ポータルが注目され，多くの企業がサービスを提供した[2]．最近では，主にコスト削減を目的としたインバウンドコールの自動応答サービスだけではなく，顧客満足度を改善する手段の1つとして，アウトバウンドコールによる情報通知サービスなどにも応用範囲が広がっており，大きなビジネスに発展している（図-1）．

●車載機器における音声インタフェース

一方，パーソナルデバイスに音声インタフェースが広く普及したのは1990年代の日本のカーナビゲーションシステムがおそらく世界でも最初の事例だろう．音声によるルートガイドは今ではほぼすべてのカーナビゲーションシステムで利用されている．また，携帯電話の普及に伴うハンズフリー音声ダイヤルの需要など，“アイズビジー・ハンズビジー”の典型的な状況として，自動車内での音声インタフェースの重要性は非常に高い．

日本のカーナビでは1990年代の後半からすでに音声認識や音声合成といった機能が，カタログスペック上の分かりやすい差別化要素として扱われ，新製品が出るたびに，認識可能な語彙のサイズや連続音声認識といった機能が競われた．その結果として，実際の使い勝手を無視したシステムが多数開発され，残念ながら多くのユーザに音声インタフェース全般に対する不信感を植え付けた面があったと筆者らは考えている．実際，2000年代前半には多くの市販ナビから音声認識機能が削除されたり，Web上のユーザ評価記事などでも機能比較の項目に含まれなくなるといった事態も起きている．実際の使い勝手

☆1 InformationWeek 2008.5.19, http://www.informationweek.com/news/software/enterpriseapps/showArticle.jhtml?articleID=207800986
☆2 日本では電話音声自動応答サービスは米国に比べるとあまり普及しなかった．

図-1　音声インタフェースの利用分野の例

や，ユーザの陥る失敗を十分精査することなく，精度を伴わないスペック上だけの機能拡張を行ったのでは，結局ユーザには受け入れられないことを我々音声技術者が改めて思い知らされた次第である．

音声インタフェースはうまく利用できれば画面注視時間を減らすことことに貢献し，結果的に安全運転につながると考えられる．しかし，単にスイッチやキーボードの代わりという理解でいると，誤認識などによって予期せぬディストラクションを招く可能性もあるので，アプリケーションの設計には注意が必要である．この観点では最近のカーナビはユーザを迷わせない工夫がしっかりしており，かなり改良が進んできたといえる．

一方，海外，特に米国では，1990年代後半からOnStarのようなテレマティックス端末等に音声認識機能が提供されていたが，離散発声の数字認識程度のものであった．カーナビ自体の普及の遅れもあったが，日本語のカーナビと同様に車載機器の音声認識インタフェース機能が大幅に拡張されたのは2003年頃以降であり，日本ではむしろ音声インタフェースに対するそれまでの期待感が薄れたころでもあった．

なお，海外系のシステムでは音声インタフェースの開発にあたっても効率の良い多言語適用を前提としており，数言語が同時開発されることも珍しくない．言語依存性はゼロにはできないので，多言語共通の基盤と，言語依存性を受け持つ部分の合理的な分離が重要である．一方で，日本語インタフェースだけは完全な独自仕様としているケースも多く，開発効率阻害要因の1つとなっていたと思われる．結果として，カーナビの音声インタフェース導入では欧米に10年近くも先行していた日本勢だが，海外展開時にはその優位性を十分活かすことができなかった．

●携帯デバイスにおける音声インタフェース

車載機器以外では，文字入力に難点のある携帯デバイスにおいても，音声インタフェースに対する期待は大きい．世界的に見ればすでに膨大な数の電話機や携帯デバイスに音声認識などの機能が組込みソフトとして搭載されている．

一方，Google，Nuance，AT&T，Microsoft/Tellme，Yahoo!/Vlingoなどがスマートフォン向けの音声検索サービスを提供して話題を集めている．これらは組込みソフトではなく，サーバベースの認識システムと考えられるが，通信速度の高速化やクラウド化されたサーバ群のおかげで，応答速度も過去のサービスに比べ格段に早くなっている．通信速度やそのコストがネックとなっていた時代には，DSR（Distributed Speech Recognition）の技術が使われていたが，今では音声程度のデータは通信上さほど問題とならなくなり，結果として携帯デバイス側に特別なプログラムを用意する必要がなくなっている．また，たとえばGoogleが提供している音声検索は英語だけではなく，すでに中国語，日本語，ドイツ語，イタリア語，フランス語，スペイン

語，韓国語などもサポートしており，徐々にその対応言語数を増やしている．これらのサービスは組込み機器向けのプログラムにありがちなリソース上の制約が問題とならないこともあって，その精度の高さと，語彙の豊富さには驚かされる．これだけ性能が高ければ，入力速度や精度の観点から見てもキーボードやテンキーに見劣りすることはなく，携帯デバイスでの検索語入力に限れば音声入力が新たな入力手段の１つとして広く受け入れられることは間違いないだろう．

一方，ディクテーションと呼ばれる音声テキスト変換技術も，世界的に見て携帯メールの利用の割合が多い日本人には特に役立つ技術である．また，欧米ではボイスメールの需要が高く，ボイスメールの音声をセンターで正しくテキスト化して電子メールやショートメッセージとして送信するサービスなども提供され始めている．過去には組込みデバイス上で動くディクテーション・ソフトウェアもいくつか開発されたが，こちらも制約の少ない，サーバベースの認識サービスが提供され始めている．米国ではいくつかのベンチャー企業のソフトに加え，Nuance が iPhone 向けの Dragon Dictation を提供している．また，Google は 2010 年 8 月に Voice Actions for Android というアプリケーションを米国でリリースした．こちらもインターネット検索に加えてメールのテキスト入力が可能になっている．このように，スマートフォン向け音声インタフェースに関する競争は日々激化しつづけている．

車載音声インタフェースに見るイノベーションの可能性

これまでに述べたように，電話音声自動応答，医療用ディクテーションといった既存のビジネス上の成功例に加えて，スマートフォンに代表される携帯デバイスでも，今後，音声インタフェースによる新たなパラダイムシフトが起こりそうな気配である．

一方，カーナビに代表される車載機器の操作についても，アイズビジー・ハンズビジー環境における音声インタフェースの重要性に揺るぎはない．ただ，他の成功例に比べると，精度を伴わないスペック上だけの機能拡張を追求した歴史があり，真に役立つインタフェースに仕上がっていない側面もある．

本章では，特に車載機器操作を例として，ユーザが直観的に使える新しい音声インタフェースに関する検討結果を紹介する．

阻害要因	詳細分類	例
雑音	定常雑音	走行雑音，エンジン騒音
	非定常雑音	音楽，対向車の通過音
	混合音声	同乗者との発話衝突，ラジオ
表現のゆらぎ	定型コマンド以外の発話	「ガソリン入れたい」
	不要語の挿入	「えっと，お台場まで」
ユーザの誤操作や過信	発声方法上の問題	発話スイッチ操作の誤り
	コマンドの覚え間違い	「外気循環」
	認識対象語以外の発話（未知語）	
	情報の欠落	「あれ見せて」

表 -1　車載機器における音声認識機能の阻害要因

●車載音声インタフェースの課題

表 -1 に，現在の車載音声インタフェースの認識性能を劣化させると考えられる要因の一部を記した．

ここに示すように，車内では走行雑音，音楽，会話といった音響的な阻害要因も多いが，特に走行雑音等については長年にわたって多くの手法が提案され，対策が施されてきた．また，ユーザが車載音声インタフェース操作時に陥る失敗への対策も大変重要であるが，適切なガイダンスを随時与えることで誤操作やコマンド以外の発話を積極的に防止するための研究も行われている[3)]．

●車載音声インタフェースの改良―拡張音声コマンド方式

ここで我々が目指したものはマニュアルを一切読まないでも直観的に操作できる車載音声インタフェースの実現である．表 -1 において，「表現のゆらぎ」に分類した阻害要因についての対策に相当する．

電話音声自動応答システムなどでは当然の前提であるが，車載機器では通常，そのようには考えられてこなかった．機器の提供側が，「ユーザがマニュ

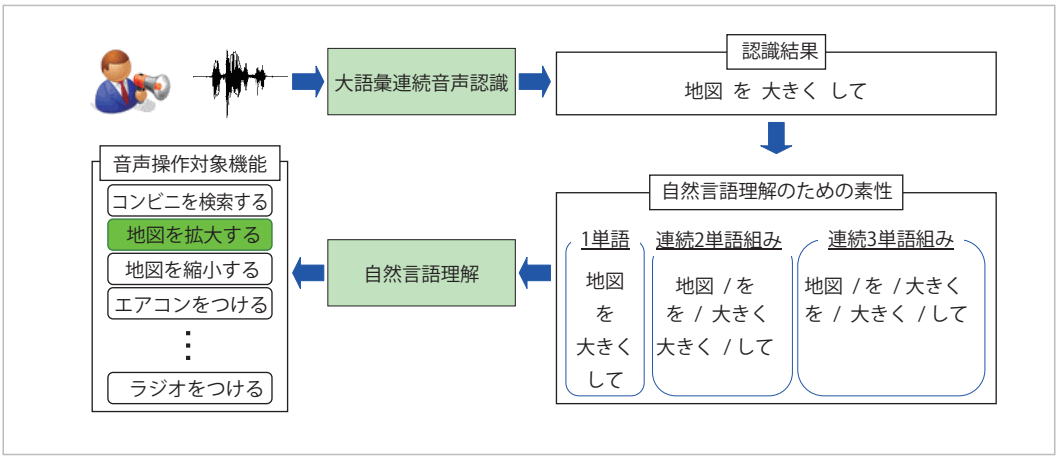

図 -2　拡張音声コマンド方式の処理の流れ

アルを読んで事前に操作方法やコマンドを覚えてくれる」と期待していた側面がある．また，多くのケースでユーザの使い勝手よりも，開発者側の都合が重視され，結果として既存のスイッチ操作を，ただ単純に音声による操作に置き換えただけのインタフェースが提供されることになった．この場合，スイッチに紐付けされた機能の名称がそのまま音声コマンドとして提供されていることが多い（音声コマンド方式）．

しかし，実際のユーザは，マニュアルを読まないことも多く，音声コマンド方式のコマンドを拡充するだけでは対応できないくらい多種多様な発話を行う．このような発話に対応するために，大語彙連続音声認識で発話を書き起こし，その結果に対して自然言語理解を行い，ユーザの意図した機能を判断する，拡張音声コマンド方式を検討した☆3, 4)．

図 -2 に拡張音声コマンド方式の処理の流れを示す．電話音声自動応答システムでは放棄呼数やオペレータ呼び出し数を減らし，また，平均対応時間を短縮するという明確な目標に向け，コールの詳細な分析や試行錯誤が繰り返され，結果的に効率の良いシステムが構築されているが，この拡張音声コマンド方式は，高度な電話音声自動応答システムですでに実用化されている音声理解技術を，車載用に転用したものと見ることもできる．

☆3 適切なガイダンスを与えることでコマンド以外の発話を防止する技術とは，解決しようとする課題は同じだが，アプローチが異なる．

拡張音声コマンド方式の効果を調べるため，ユーザの車載機器操作のための発話について，表 -2 に示した 5 段階のレベル（1 から 5）と 2 種類の扱うことができないレベル（A : Ambiguous, B : Bad）を定義した．各々のレベルについて，音声コマンドを文法で記述した音声コマンド方式で対応できるかどうかを左から 3 列目に，また，拡張音声コマンド方式での大語彙連続音声認識，自然言語理解が可能かどうかを 4，5 列目に，“○”，“△”，“×”で示した．たとえば，音声コマンド方式ではレベル 1 には対応でき，文法を豊富にすることによりレベル 2 にもある程度対応できるが，それ以上のレベルについては対応できないことを示している．拡張音声コマンド方式の場合，レベル 4 以降では，認識対象外の単語が出現して認識できない場合があるため，大語彙連続音声認識は“△”とした．自然言語理解についても，レベル 4 では，認識できなかった単語の影響で誤りが増大する可能性があるため，“△”としている．レベル 5，A，B については，理解に高度な背景知識を利用する必要があるなど，入力された一文の発話だけでは正しい判断ができない．このように，拡張音声コマンド方式を利用した場合には，レベル 3 までは対応が可能であり，レベル 4 についても一部の発話には対応できる可能性がある．

ユーザの発話を大量に収集し，それを分類した結果，レベル 1 に含まれる発話の割合は 10% 以下であった．それに対してレベル 1 から 4 までを含めると，その割合は 95% を超えていた．これは，拡

	発話内容	音声コマンド	拡張音声コマンド		例
			認識	理解	
1	コマンド	○	○	○	近くのコンビニ
2	事前に想定できる言い回し	△	○	○	近くのコンビニを検索
3	事前に準備できる言い回しではないが意味が十分伝わる	×	○	○	近隣コンビニ検索開始してください コンビニで買い物したい
4	冗長な表現や認識対象外の未知語を含む	×	△	△	行列のできるレストランでシチューを食べたいなぁ コンビニまでコーラを買いに行く
5	理解に高度な背景知識を必要とする発話	×	△	×	いつもの寿司 あの時のイタリアン
Ambiguous	他の機能と区別できない	×	△	×	ちょっと下げて（「温度」か「音量」か分からない）
Bad	内容・意味が不明	×	△	×	あれいいね

表 -2　ユーザの発話の分類と拡張音声コマンド方式の優位性

張音声コマンド方式によりユーザの多種多様な発話を処理できることを示唆している．

また，実際に評価用に音声データを収集し，マニュアルに基づく音声コマンド方式と，ここで紹介した拡張音声コマンド方式の性能を比較した．自動車内でのエアコン操作に関する発話を対象とし，総発話数に対して，意図した機能を起動することができた発話数の割合を現すタスク達成率で評価を行った．音声コマンド方式では 23.0% であったタスク達成率を，拡張音声コマンド方式により 95.0% にまで改善することができた．

図 -3 に示すように，拡張音声コマンド方式によって，音声コマンド方式では被覆することができなかった表現のゆらぎの多くを被覆することが可能となると考えている．

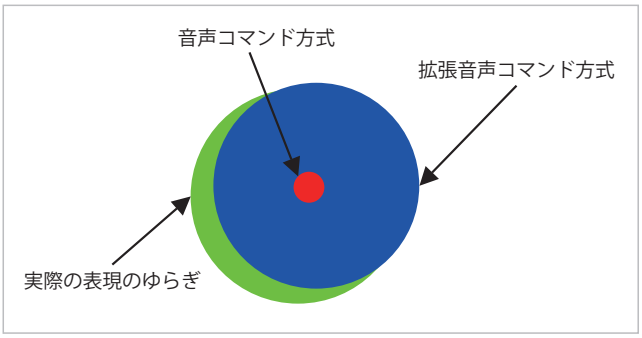

図 -3　拡張音声コマンド方式で被覆できる発話の範囲のイメージ

さらなるイノベーションの可能性

ここまで，車載用機器，主に車載型のカーナビゲーションの，音声インタフェースの使い勝手を改善する 1 つの方法について紹介した．ただ，ある程度規模の大きな辞書を用いて，自由な語順の発話を受理する必要があり，これらを組込み機器の限られたリソース上で実現するのは決して容易ではなかった．

一方，日本では車載型カーナビの需要がまだ多いが，世界的な出荷台数で見れば PND（Personal Navigation Device）と呼ばれる簡易型カーナビの利用が圧倒的に多くなっているという事実がある．

PND は車載型のカーナビに比べ廉価なデバイスとの位置づけの製品なので，音声インタフェースもこれまで以上に限られたリソースで，機能を絞って実現されることが多く，先に紹介したような高度な音声理解技術の導入はさらに困難である．ただ，最近ではその PND ですら，将来的にはスマートフォンなどのナビアプリケーションで多くが置き換えられてしまうだろうという予測がある．

ここで注意すべきことは，カーナビやテレマティックス端末，あるいはスマートフォンへの応用に限らず，安価で高速な通信が常時可能になれば，どのようなデバイスであっても，前述の高速通信とクラウドに支えられたサーバベースの音声インタフェース技術によって，リソースの問題を一気に解決できる可能性があるという点である．言い換えれば，あらゆる場面，あらゆるデバイスで潤沢な計算リソースと，最新の情報を活用し，現在の最高の音声インタフェースを安価にユーザに提供できる可能性が出てきているのである（図 -4）．たとえば，テレビなどの家電やゲーム機など，過去に音声インタフェースの適用例として注目されたデバイスは，すでにネッ

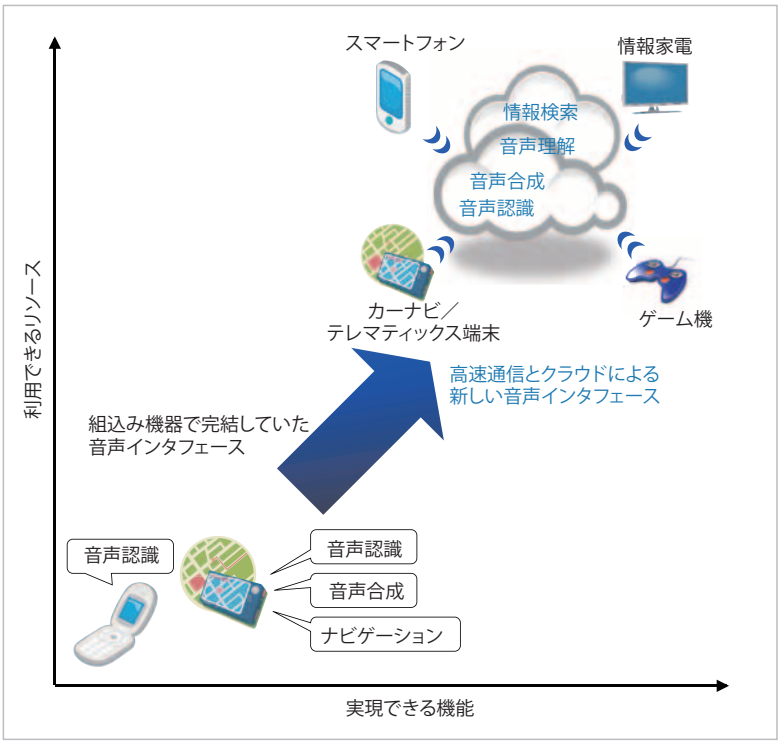

図-4　パラダイムシフト（高速通信とクラウドによりインタフェースの世界が変わる）

トワークに接続されていることもあり，改めて応用先として見直されるだろう．

また，対象は音声認識に限らない．音声合成も，サーバベースとなれば，高品質で多彩な声の再生が可能となるし，ネットワークの先には膨大な情報資源がある．

残された課題は通信に伴う応答速度の低下や，高速通信が困難な状況への対応であるが，応答速度を要求される場面や通信が困難な状況では組込み型の音声認識を実装し，それ以外はサーバベースの音声認識を利用するハイブリッド型のシステムを実現すればよい．

また，サーバベースの音声認識は別の観点でも革新をもたらす．それは，ユーザの発話，利用状況といった大量の情報をサービス提供者が逐次入手できるということである．すでにWebの世界ではGoogleが検索情報を一手に入手することで大きなビジネスを生み出しているが，音声インタフェース利用時にも同様に，ビジネス上大いに価値のある情報を収集できることになる．もちろん，収集した音声データを用いて音声認識精度の改善や，音声インタフェースとしての性能改善を効率的に進めることができることは言うまでもない．

すでに音声認識や合成の基本性能もこの10年で大幅に改善されている．これらの機能を活用し，今後，すばらしい音声アプリケーションが続々と登場することを期待している．

参考文献

1) Kuo, H. -K. and Lee, C. -H. : Discriminative Training of Natural Language Call Routers, IEEE Transactions on Speech and Audio Processing, Vol.11, No.1, pp.24-35 (2003).
2) Bacchiani, M., Beaufays, F., Schalkwyk, J., Schuster, M. and Strope, B. : Deploying GOOG-411 : Early Lessons in Data, Measurement, and Testing, In Proc. ICASSP, pp.5260-5263 (2008).
3) 岡本　淳，庄境　誠：人間の多様な振る舞いを考慮した音声UIの必要性，情処学研報，2009-SLP-78-10 (2009).
4) 倉田岳人，市川　治，西村雅史：ユーザの発話傾向分析に基づく車載機器操作のための音声入力手法の検討，電子情報通信学会論文誌，Vol.J93-D, No.10, pp.2107-2117 (2010).

（平成22年8月31日受付）

西村 雅史（正会員）nisimura@jp.ibm.com
1983年大阪大学大学院基礎工学研究科物理系博士前期課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所にて，音声認識などの音声言語情報処理の研究に従事．同社主席研究員．工学博士．1998年本会山下記念研究賞，1999年日本音響学会技術開発賞各受賞．IEEE，電子情報通信学会，日本音響学会各会員．

倉田 岳人（正会員）gakuto@jp.ibm.com
2004年東京大学大学院情報理工学系研究科電子情報学専攻修士課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所にて，音声認識などの音声言語情報処理の研究に従事．同社主任研究員．日本音響学会会員．