社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74 (19)
2008 / 12 / 10

[招待講演] 音声認識応用システム開発の新パラダイム

小林 哲則†

†早稲田大学 理工学術院
〒169-8555 東京都新宿区大久保 3-4-1

あらまし 従来音声認識応用システムの開発は，エンジン開発者から提供を受けたエンジンをアプリ開発者がシステムに組み込みユーザに渡すという，一方向的開発パラダイムに沿って行われてきた。ここで，アプリ開発者とエンジン開発者が密なる連携をとれない場合，十分に使い勝手の良い音声応用システムはできない。ここでは，エンジン開発者とアプリ開発者，さらにはユーザも含めた密なる連携の下に音声応用システムを開発することの重要性を述べ，それを実現するためのフレームワークを紹介する。また，これを基礎とした今後の音声認識応用システム開発のいくつかの形について述べる。

キーワード 音声認識，音声認識応用，ソフトウェア工学，プロキシエージェント，

A New Paradigm for Speech Application System Development

Tetsunori KOBAYASHI†

†Faculty of Science and Engineering, Waseda University,
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan

Abstract Speech recognition application systems have been developed along so called one-directional development paradigm: the engine builders develop engines and pass them to the application programmers, application programmers develop applications with the engines and pass them to users, and the users just use the system. When the application programmers and the engine builders cannot take close cooperation, and this is often the case, they cannot make convenient speech recognition application systems. Here, we introduce a framework for the development of speech recognition application systems which forms close cooperation among engine builders, application programmers and also users. We also describe some desirable future of the speech recognition application system development based on the framework.

Keyword Speech recognition, Speech recognition application, Software engineering, Proxy agent

1. はじめに

音声認識システムの性能は著しく向上した。NHK のニュース字幕変換器は既に数年前からほぼ完璧な書き起こしを実現している[1]。最近の京都大学の国会中継の議事録作成器のデモを見ても，誤り個所を見つけるのに苦労するほどである[2]。

一方で，我々が音声認識器のユーザとしてその恩恵にあずかれる場面は極端に少ない。私の車の音声カーナビはと言えば簡単なコマンドすら受け付けてくれない。ある日のこと，カーナビの案内に従って運転をしていた筆者は，その案内を終わらせたかった。「ナビ終了」「ナビゲーション終了」「案内終了」いくつかの言葉を試したが，何を話してもカーナビの道案内が終わることはなかった。ときにとんでもない場所の地図を示し，ときにオーディオの設定を変え，およそナビゲーションの終了とは無関係な動作を続けた。そんなこ

とが続く中で，筆者はいつしかカーナビの音声機能を使うことを諦めた。筆者が音声認識に強い愛着を持つにも関わらず，である。

この落差はどこから来るのであろうか。カーナビには最新の技術を導入できないのであろうか。使用環境の違いであろうか。個人毎の声が多様にすぎるのであろうか。その答えは，半分はYESであり，半分はNOであろう。それらの要因は原因のひとつではありえるが，決定的な要因とは考えにくい。

また，いまだにカーナビ以外の身近なアプリケーションがないのは何故だろうか。ハンズフリーの状況でしか音声は役立たないのであろうか。筆者が音声認識の研究者であるが故のひいき目の分を割り引いてみても，音声認識の価値ある利用場面は他にも数多くあるように思われてならない。では何が必要か。筆者は，この5年間いくつかの官産学共同プロジェクトおよびその準備を通じて，音声認識エンジンの主要メーカの方々と，音声認識の実用化を阻む諸問題の整理とその解決の在り方について議論を重ねてきた。本稿では，その議論と議論を踏まえて開発したシステムを紹介するとともに，今後の音声認識応用システムの開発の形について提言をしてみたい。

2. 現状

2.1. 一方向性の開発パラダイム

現状，音声認識応用システムは，音声認識器の性能を十分に引き出していないようだ。先のカーナビの例では，「誘導中止」とさえ発話すれば案内は終了する。システムが案内を終了するために待ち受けていたその言葉を，筆者が思い浮かべることはできなかっただけの話で，認識器に罪はない。正しく待ち受け語彙を発話することができさえすれば認識ミスを起こすことはほとんどない。しかし，何という表現で待ちうけているかがユーザに見えないことは，インタフェースの透過性に係る本質的な問題[3][4]であって，システム全体としてみたときには重罪を犯していることになる。しかも，カーナビの作りをいろいろと見ると，待ち受け語彙を分かり易くしなければいけないという極基本的な事柄さえ意識していないのではないか思えてくる。また，ユーザがどのような実用の場面でどのような不都合に直面しているかをアプリケーションの開発者にフィードバックする仕組みがないことも，問題を極端に難しくしている。

このような現状には，日本における音声認識応用システムの開発体制が少なからず問題となっているようだ。日本においては，極端な分業体制の中で行われることが多い。エンジン技術者が開発したエンジンをアプリ開発者が受けとってこれにアプリをかぶせてアプリケーションを作る。情報の流れは，システムの流れと同様に，エンジン開発者からアプリ開発者，ユーザへと概ね一方通行であって，逆の流れは稀である[5](図1参照)。ここで驚くべきことであるが，アプリ開発者は，エンジンについての詳しい特性も，音声認識応用システムを利用したときのユーザの振る舞いに関する知見も持っていないことが圧倒的に多いという。

性能が上がったとは言え，音声認識器はどのように使っても100%の性能を発揮できるほどの完成度を持った部品にはなりきっていない。アプリ開発者は，エンジンの性能を引き出すための術を知る必要があるのだが，それを知る構造がないようだ。エンジンメーカ側も，アプリ開発に必要な性能情報を提示する必要があるのだが，それが十分でない。このため，どういう条件で使うとどういうことになるのか，どの程度の性能になるのかがアプリ開発側に見えていない。結果，適切なエンジンの使い方ができない，良いアプリケーションができない，音声認識が広まらない，ユーザが慣れない，性能が上がらない（性能が出るためにはある程度の慣れが必要である），とネガティブスパイラルができあがる。

また，アプリ開発者は，真の意味で問題を掴んでい

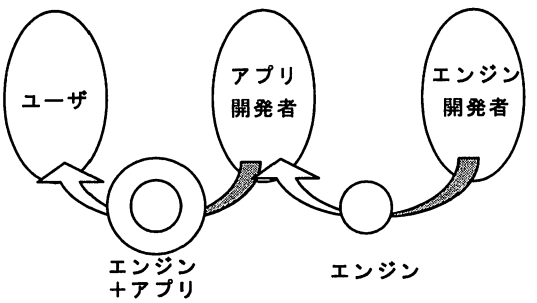

図1. 一方向性開発パラダイム

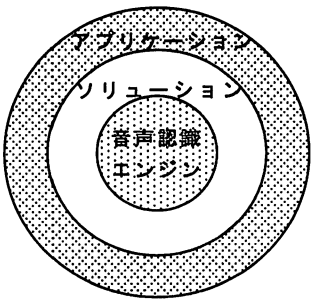

図2. 音声認識応用システムを作る技術階層

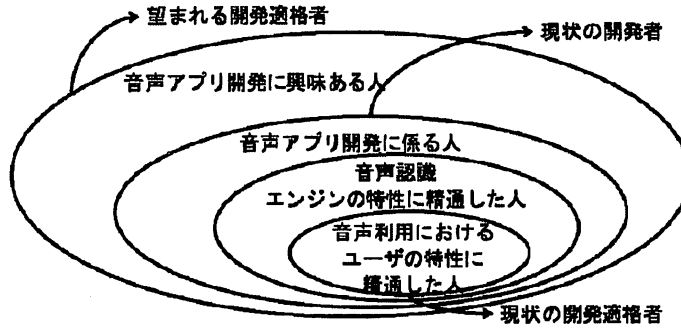

図 3. 音声認識応用システムの開発者の階層

ないという問題もある。現状で，売ったシステムの動作解析はできない。現場で何が起こっているかわからない。結果としてユーザの声が開発にフィードバックされない，よって使いやすくならない。ここでもネガティブなスパイラルが形成されてしまう。

2.2. 音声ソリューションを担う技術者の必要性

良質な音声認識応用システムができるためには，音声認識器を使いこなして良いアプリケーションにつなげるためのノウハウなり技術が必要である。音声認識器の特性，および音声利用時のユーザの特性に関する深い知識を持った上で，認識器をカスタマイズし，応用システムを設計することが必要なのである。実際，音声認識のビジネスをうまく軌道に乗せているいくつかの企業は，このエンジンとアプリケーションとのつなぎの技術の蓄積に多くの人材とコストをかけている。しかし、その効果をより顕著なものとするためには、今一歩スケールメリットを出せるための組織なり技術なりが必要とされているように思われる。

ここでは，この音声認識器とアプリケーションとのつなぐ役割を音声ソリューションと呼ぶことにする(図 2 参照)。音声技術を持って，有用なアプリをどのように実現するかについての解を与える役割である。誰かが，音声ソリューションの役割を担わなければ，使い勝手の良いシステムはできない。では，誰がこの役割を負うべきであろうか。

この役割をエンジン開発側に求めることは，日本のエンジンメーカが置かれた状況からすると，現実的ではない。膨大な数のアプリに万能な認識器を開発することは困難であって，どこかに特化して強みを発揮するエンジンを開発している状況を考えると，エンジンメーカが音声ソリューションをビジネスの軸として展開することが難しいことは容易に想像できる。日本のエンジンメーカが一般に大企業であって，細かなビジネスを相手にできないことも，音声ソリューションを手掛けることを難しくしているようにも思われる。(勿論，やっていけないわけではない。いくつかの企業がこれに挑戦し，成果を挙げていることもまた事実である。)

また，アプリ開発者がこの役割を担うために，音声認識エンジンと音声利用時のユーザファクタの専門家であらねばならないというならば，これも筋の良い話ではない。音声の専門家であることを求められるなら，音声アプリの開発者層は増えそうもない。そうなると，音声アプリは決して広がることはない。良いアプリが増えるためには音声アプリケーションの作り手が増えることが絶対的な条件なのである(図 3 参照)。

3. 音声認識基盤技術の研究開発プロジェクト

3.1. 双方向性開発パラダイム

上記現状把握に基づいて、2006 年に発足したのが経済産業省 戦略的技術開発委託による「情報家電センサー・ヒューマンインターフェイスデバイス活用技術開発『音声認識基盤技術の研究開発』」である。本プロジェクトは，早稲田大学が受託し，東工大，旭化成，ＮＥＣ，沖電気，東芝，日立，三菱に再委託する形をとり，2009 年 3 月までの予定で進められている[6]。

プロジェクトでは，

① ポータビリティの高い手離れの良い音声認識技術
② 一定水準の音声インタフェースを開発するツール
③ ヒューマンファクタおよび音声アプリ開発に係る知見の集約・共有
④ システム利用時のユーザの生のデータが低コストでの蓄積

等の開発によって，音声認識応用システムの開発における音声ソリューションの技術的支援をすることを柱のひとつとした。(その他,実環境音声認識の精度向上，WFST による柔軟・高精度なデコーダ技術の開発なども柱となっている)。ユーザ・アプリ開発者・エンジン

開発者間で互いに情報を共有しながら双方向の密なる連携を実現し，継続的に育てることができる音声認識応用システムの開発環境を実現することを目指している。このことによって，アプリ開発者がエンジンやヒューマンファクタに関する深い知識がなくとも，一定水準の音声認識応用システムを開発することを可能にすることが期待される。

3.2. プロジェクトが実現した世界

中核を担うのはプロキシエージェントと呼ぶ新たな音声認識応用システムの構成要素である。プロキシエージェントは，音声認識器の開発者およびアプリ開発者の負担を抑えた上で，様々なサーバ連携を可能にする[7]。アプリとエンジンの通信を仲介しながら，必要に応じてデータとプログラムのアップロードとダウンロードを行う。サーバとの連携機能をエンジン・アプリケーション非依存に行うことを可能にすることによって，情報共有のスケールメリットを実現することを目指している。現状では，Sphinx4，VORERO，Juliusに加え，プロジェクトで東工大が開発したT^3[8]の４つのデコーダがプロキシエージェント対応になっている。

インターネット上には，プロキシエージェントと連携して音声認識応用システムに様々な拡張機能を与えるサーバ群や，開発者に有用な情報を提供するサーバ群を配した。

利用ログ蓄積サーバは，音声応用システム利用時におけるユーザの生の振る舞いをアップして蓄える。また，ツール群も用意して利用ログ解析を支援する[7]。アプリ開発者は，これらの道具立てを利用して，システムをユーザに提供した後も，システムの改良を継続的に行うことができ，またその改良結果をプロキシエージェント連携による配信機能によって，ユーザに届けることができる。

語彙情報共有サーバは，WEB 上で利用可能な情報を利用して，音声認識器のための語彙情報を提供する[9]。単語に付与したタグの集合によって語彙を管理する機能を持ち，プロキシエージェントの配信機能によって，WEB に現れる新出単語についても利用者に届けることができる。NEC が提供する略語読み付与サーバと連携して，略語の扱いも可能にする。

旭化成が開発した性能予測サーバは，標準評価データを用いてデコーダの評価を行うことができる。近い将来，少数の評価サンプルを与えることでデコーダの性能予測分布を与えることを検討している。この機能により，アプリ開発者は，ターゲットとなる環境において，どの程度の認識性能が見込まれるのを知った上でシステムの設計が可能となる。

三菱電機が開発した音声インタフェース開発ツールは，定められた表記法に沿って実現したい機能を記述すると，理にかなった音声インタフェースが実現できる仕組みとなっている。経験の少ない技術者でも，勘所を外さずに音声インタフェースが実現できる。

その他，開発の知見をパターンランゲージの形で表現し，これを開発者間で共有する仕組みも実現されている。音声認識応用システムの開発において直面する代表的な問題とその解決方法を記した手引書を，多くの技術者が共同して作ることになる。これにより，経験の少ない技術者でも，先人の失敗を繰り返すことなく効率的にシステムを開発することができる。

以上のようなサーバ群との連携に基づいて，音声認

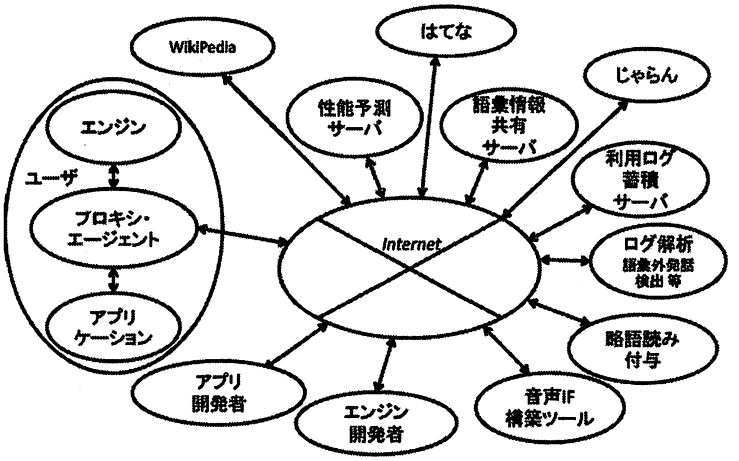

図 3. ユーザ・アプリ開発者・エンジン開発者間の連携に基づく双方向的開発パラダイム：経済産業省「音声認識基盤技術の研究開発」プロジェクトにおけるプロキシエージェントを核とするサーバ連携の形

識応用システムを開発・運用することで，開発の弱点であった音声ソリューションがカバーされ，良質の音声認識応用システムが実現されることが期待される。

もちろん，開発支援に必要な機能はプロジェクトで開発できたものの他にも数多くあって，それらについては今後地道に開発を続ける必要がある。プロキシエージェントは，機能拡張を容易に行えることを特徴としているため，新たなサーバが開発されたときも，その効果を簡単にアプリ側に伝えることが可能になっている。

4. 音声認識応用システム開発の今後

4.1. 音声ソリューションビジネスの可能性

前章に述べた開発環境の実現を受けて今後の展開であるが，この環境を利用しながら音声ソリューションを軸としてビジネス展開を図る企業が育つことがひとつの望まれる形として考えられる。

本来であれば，音声ソリューションを手掛けるべきは，エンジン開発メーカともアプリ開発メーカとも独立した専門家の集りであることが望ましい。しがらみの無いアプリ選択環境と，しがらみの無いエンジン選択環境を持って，ソリューションを提供できることが良質のアプリケーションを実現するための重要な要件だからだ。前章で紹介した経済産業省のプロジェクトの成果は，アプリ・エンジン非依存に利用できる形で提供されている。このことは，音声ソリューションに携わる企業が，業務を効率化し，スケールメリットを出していくために極めて重要な意味を持つものと考える。

現状のエンジンメーカから，カーブアウトするなどして音声技術者が結集し，音声ソリューションに関するビジネス展開をするようになれば，日本においても優れた音声応用システムが巷にあふれるかもしれない。

4.2. アマチュアによる音声ソリューション

前節に述べたようにソリューションのプロ組織が育つことを期待する一方で，エンジンメーカがそれぞれに力を持つ日本においては音声ソリューションを担う力のある企業が育つのは難しいのかもしれないという思いもある。

筆者が今希望を持っているのは，むしろアマチュアの台頭である。Wikipedia が Encarta に勝った(?)ように，もしかしたら，真にユーザの心を掴む音声認識応用システムは，少数のプロの力によってではなく，アマチュア大連合のチーム力によって実現されるかもしれない。

興味深いのは，産総研・後藤，緒方による一般利用者の間接的音声認識器開発参加の枠組み[10]と，和歌山大・西村や早大・中野による WEB アプリ上での音声認識器のマッシュアップの枠組み[11][12]である。後藤らのシステムは，PodCast の音声検索と絡めて，音声の書き起こしに一般利用者を巻き込むものである。一般利用者に書き起こしを手伝う動機が本当にあるのかという心配をよそに，安定した利用者を獲得できている。西村や中野の枠組みは，容易なアプリ開発を可能にすることで，音声認識開発のプレーヤーを増やし，良質なアプリが開発される可能性を増やす。適切な仕組みを持って「量」を操ることに成功すれば，それはいつしか「質」の変化を生む。経産のプロジェクトで開発した技術や，ユーザ連携を目指す様々なアプローチが絡み合い，ポジティブなスパイラルを形成するなら，愉快なことである。

5. むすび

本稿では，音声認識応用システムの新たな開発パラダイムとして，ユーザ・アプリ開発者・エンジン開発者の有機的な連携に基づいてシステム開発を行う方法論について述べた。この方法論は，継続的なシステム開発と修正システムの再配信の枠組みを持つもので，音声ソリューションにコストをかけられない体制においても，良質なシステムを開発できる可能性があることを述べた。

この継続的な開発と再配信によって商品の質を高める方法論は，ソフトウェアの世界では極普通のことである。しかしながら，ソフトウェア以外の世界ではなかなかこれを受け入れる素地がないようだ。「我社がユーザの手を借りなければ完成に至らない未熟なシステムを市場に出すなどまかりならん」という発想である。未熟なシステムでも，成熟への道筋をつけた上での市場にでるなら問題は少なかろうが，そういった考えが受け入れられないことが多いのは若干残念である。

また，通常であれば音声のようにヒューマンファクタが使い勝手に深く影響を持つシステムの普及には，多かれ少なかれデファクトをとるシステムの存在が必要である。システムがユーザに受け入れられるためには，ユーザがある程度システムに慣れる必要があり，このためにはインタフェースにアプリを超えた一貫性が求められる（どのアプリを使ってもおおよそ同じ使い方ができることが望まれる）。通常，それを主導するのはデファクトをとるシステムである。しかしながら，音声認識においては，なかなかデファクトをとるシステムが実現しない。現状，エンジン，ソリューション，アプリケーションとバランス良く開発できる体制がなく，デファクトをとるほどに優良なシステムが作りにくいからと考える。

一社であるいは一つのアプリでデフォルトをとるほどに優れたシステムを開発できないならば，業界全

体でインタフェースに一貫性をもたしてはどうかと思うのであるが，これもまた結構綱引きがあって難しいようだ。であれば，せめて開発知見の共有を進めてほしいと願っている。主観的な主張に終始すれば，なかなか綱引きは終わらない。客観的なデータの分析は我々の進むべき方向を教えてくれるに違いない。共同で音声認識の価値を周知させ，パイを広げることこそが重要と考える。認識器の使い方さえ適切であれば，我々がユーザとしてその恩恵にあずかる場面が少ないはずがない。

文　献

[1] 今井亨，伊藤崇之，“放送における音声認識の取り組みと課題，” NHK 技研　R&D, No.89, Jan. 2005.

[2] 河原達也，“国会審議及び大学講義の自動音声認識，” JEITA シンポジウム，Speech Technology: Today and Tomorrow, Oct.2008.

[3] 西本 卓也,志田 修利,小林 哲則,白井 克彦，"マルチモーダル入力環境下に おける音声の協調的利用 ---音声作図システム S-tgif の設計と評価---," 電子情報通信学会誌　DII, J78-D-II, 12, pp.2176-2183, Dec. 1996.

[4] D.A.ノーマン,(野島久雄訳) “誰のためのデザイン?—認知科学者のデザイン原論，” 新曜社認知科学選書.

[5] 2005 年度 新エネルギー・産業技術総合開発機構 音声認識技術実用化に向けた先導研究事業「音声認識技術実用化に向けた先導研究」報告書

[6] 2006 年度 経済産業省 戦略的技術開発委託費 情報家電センサー・ヒューマンインターフェイスデバイス活用技術開発「音声認識基盤技術の研究開発」報告書

[7] Teppei Nakano, Shinya Fujie, Tetsunori Kobayashi,"Extensible speech recognition system using Proxy-Agent," Proc. IEEE ASRU 2007, Dec. 2007.

[8] 大西翼，ディクソン ポール，古井 貞熙，"WFST 音声認識デコーダの開発とその性能評価，" 情報処理学会研究報告，音声言語情報処理，2007-SLP-68(1), Oct. 2007.

[9] 中野 鐵兵，佐々木 浩，藤江 真也，小林 哲則,"WWW を用いた語彙情報の収集・共有・管理システム，" 情報処理学会　音声言語情報処理研究会, SIG-SLP-71-12, May 2008.

[10] Masataka Goto, Jun Ogata, and Kouichirou Eto: PodCastle: A Web 2.0 Approach to Speech Recognition Research, Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007), pp.2397-2400, August 2007.

[11] Ryuichi Nisimura, Jumpei Miyake, Hideki Kawahara, Toshio Irino, "Speech-to-text input method for Web system using Javascript", Proc. 2008 IEEE Workshop on Spoken Language Technology (SLT2008), Dec. 2008. (to appear)

[12] 中野鐵兵，藤江 真也，小林 哲則，"Proxy-Agent を用いた音声認識対応ウェブアプリケーション開発フレームワークの提案と実装，" 情報処理学会研究報告, 2008-SLP-70-15, 2007