PodCastle: 集合知を利用した新たな研究アプローチ「音声認識研究 2.0」*

○後藤 真孝，緒方 淳，江渡 浩一郎 (産総研)

1　はじめに

本招待講演では，Web 2.0[1] に基づく Web サービスを提供し，音声認識誤りを含む認識結果を幅広く開示することで，不特定多数のユーザの協力を得て音声認識技術を発展させていく研究アプローチ**「音声認識研究 2.0」**[2),3)] を紹介する．我々は，これを具現化した音声認識のキラーアプリケーションを目指して，音声認識に基づくポッドキャスト検索・閲覧用 Web サービス**「PodCastle」（ポッドキャッスル）**の公開を開始した[4)]．PodCastle では，ユーザが Web 上の日本語のポッドキャストを全文検索できるだけでなく，認識結果の全文テキストも閲覧でき，さらに誤認識箇所を容易に訂正することもできる．これにより，ユーザが利用しながら訂正すると認識性能と検索性能が向上し，さらなる利用が促せるというポジティブスパイラルが生じることが期待できる．

2　本研究の二つの目的

多くのエンドユーザ（音声認識利用者）は，音声認識が有用な技術であることをまだ実感していない．音声認識研究者は，音声認識が高度な技術に基づいており，どのような音声が認識しやすく高い性能を示すかを知っている．一方，エンドユーザは音声認識の原理を知らず，どのような音声が認識されやすいかは充分には理解していない．そのため，過去に自分の音声が正しく認識されなかった経験等があると，そのときの印象で有効性に疑問を抱き，使わなくなることが多い．様々な研究により認識率が向上し，文献 5) で「なぜ音声認識は使われないか」が分析されたときの状況から進展してはいるものの[6)]，ユーザの利用率が低いという問題は依然として解決されていない[7)~9)]．

本研究の第一の目的は，この問題を解決すべく，ユーザに現在の音声認識の技術レベルを把握してもらい，その普及と実用化を促進することにある．そこで我々は，ユーザがポッドキャストを検索・閲覧できる Web サービス「PodCastle」を公開し，様々な音声の認識結果の全文テキストをユーザと共有することを可能にする．ポッドキャストは，Web 上の音声データとして多数公開されているため，ユーザ自身が発声しなくても，様々な難易度の音声に対する認識結果を閲覧することで，認識技術の現状が把握できる．例えば，マイク入力した自分の音声が誤認識されると，それを不快あるいは恥ずかしいと思うユーザがいるが，既に公開されているポッドキャストの認識結果を見てもそうした問題がなく，利用に躊躇がない．

(i) ユーザが音声認識を体験することで，その性能を理解する．
(ii) 音声認識の性能向上にユーザが貢献する．
(iii) 性能が向上したら，それがより良いユーザ体験に結びつく．

図 1: 「利用される音声認識」へ向けたポジティブスパイラル（(i)〜(iii) の各段階が繰り返される好循環）

しかし，ポッドキャストの内容や収録環境は多種多様であり，現在の音声認識技術では多くの誤認識箇所が発生する．こうした問題に対する典型的アプローチは，認識対象の音声データを大量に収録してコーパスを作成し，書き起こしテキストを用意して学習・適応する方法である．ただし，このアプローチでポッドキャストの全文検索を実現しようとすると，あらゆる音声に対するコーパスを整備する状況に近くなり，コストや労力の観点からも現実的でない．

本研究の第二の目的は，この問題を解決すべく，事前に対象となるコーパスを用意する考えを捨て，不特定多数のユーザの力を借りて音声情報検索と音声認識の性能向上を実現することにある．PodCastle では，音声認識技術では不可避な誤認識箇所をユーザに訂正する協力をしてもらうことで，適切に検索できるようにしていく．さらに，その訂正履歴を学習に利用することで，運用中に自動的に音声認識の性能向上が図れる仕組みを実現する．これは，ユーザに「音声認識を育ててもらう」アプローチと言える．

3　「音声認識研究 2.0」とは？

文献 2),3) では，このように「ユーザに対して音声認識の現状を積極的に開示し，ユーザの協力を得て音声認識技術を発展させていく研究アプローチ」を「音声認識研究 2.0」と名付けた．これにより，研究分野全体での問題意識の共有を図り，問題解決へ向けて力を合わせて取り組んでいけることを狙っている．これは，Web 2.0[1] を意識して付けた名称であり，Web 2.0 の特長を取り入れて不特定多数のエンドユーザの協力を仰ぐことで，音声認識の性能向上と実用化（利用率の向上）を共に実現していくことを目指す．

その実現のためには，図 1 のポジティブスパイラルを回すことが重要だが，従来はこの三段階のそれぞれに以下に述べるような阻害要因があった．

- (i) の性能理解に関しては，従来は，ユーザ自身の発声を認識した結果を見て，性能を誤解する可能性が高かった．多くの音声認識研究者は，他人の適切な発声（コーパス中の音声）を認識した結果を

*PodCastle: A New Research Approach “Speech Recognition Research 2.0” Using Wisdom of Crowds. by Masataka Goto, Jun Ogata, and Kouichirou Eto (AIST)

表 1: 従来の音声認識研究のアプローチ「音声認識研究 1.0」と本研究で提案するアプローチ「音声認識研究 2.0」の対比

音声認識研究 1.0	音声認識研究 2.0
スタンドアロンアプリ	Web サービス
ディクテーション	検索・閲覧
コーパス	Web 上のデータ
話題限定	話題非限定
書き起こし	アノテーション
未知語	未アノテーション語
専門家参加	ユーザ参加
個人的訂正	社会的訂正
個人知	集合知
完成版	永久にベータ版

上記は，Web 1.0 と Web 2.0 を対比した文献 1) の表に影響を受けて記述した．これらの項目を満たすほど音声認識研究 2.0 的な研究事例と言えるが，Web 2.0 の場合と同様に，すべてを満たさなければならないわけではない．

目にする機会が多いため，性能を誤解することはなかった．しかし，ユーザは何度か自分の発声が認識されない体験をするだけで，他の人の音声も同様に認識されないものだと誤解することがあった．

- (ii) の性能向上に関しては，従来，話者適応のためにユーザに例文を発声させたり，未知語を辞書登録させたりすることが多かった．しかし，そうしたエンドユーザによる性能改善が，他のユーザと共有されて再利用されることはなく，総体としての音声認識の性能向上には，音声認識研究者しか貢献できなかった．そのために，不特定多数のユーザが共に性能向上を実感して，それに共同で貢献していくことを動機付ける要因はなかった．
- (iii) のユーザ体験向上に関しては，音声認識研究者の手元で日々性能が向上していても，その高い性能をユーザが体験する機会は限られていた．音声認識を利用した市販ソフトウェアでも，数ヶ月〜数年のバージョンアップのサイクルでしかユーザは性能向上を体験できなかった．

音声認識研究 2.0 では，これらを解決することで，図 1 のポジティブスパイラルを回し，音声認識を取り巻く状況を変革することを目指す．従来の典型的な研究アプローチ（以下,「音声認識研究 1.0」と呼ぶ）との対比を表 1 に示す．ここでは対比する便宜上，従来の研究アプローチを「音声認識研究 1.0」と名付けたが，それは決して劣るものでも不要なものでもなく，今後の音声認識の発展のために継続して研究することが必要不可欠であることは間違いない☆1．これは,「音声認識研究 1.0」を土台として，それに加えて「音声認識研究 2.0」のアプローチにも取り組むべきであるという提案である．なお，音声認識の手法自体について議論しているのではなく，研究の方法論，アプローチについて議論しているため,「音声認識 2.0」ではなく「音声認識研究 2.0」と名付けた．

以下，表 1 の項目について説明しながら，図 1 がどのように実現されるかを述べる．

- 音声認識研究 2.0 では，コーパスに基づいて学習した音声認識システムをディクテーション等のスタンドアロンアプリケーションとして提供するのではなく，**Web 上の音声データ**を対象に，ユーザが直接**検索・閲覧**できる **Web サービス**を実現する．これにより，図 1(i) の性能理解が促進される．
- しかし，Web 上の音声データを対象とすると，話題が従来の音声認識研究のようには限定できず，コーパスやその書き起こしも整備されていないため，多くの誤認識が起き，未知語も多くなる．そこで音声認識研究 2.0 では，**話題非限定**な状況で多様な音声データの認識に挑戦し，誤認識箇所はユーザに訂正してもらって検索可能にする方針をとる．つまり，各音声データの検索用**アノテーション**として，書き起こしに相当する全文テキストをユーザの協力により整備していく．ここで重要なのは，その訂正内容を学習することで，まだ訂正していない部分や他の音声データに対する認識結果が改善される点である☆2．未知語に関しても，ユーザがまだ訂正していない**未アノテーション語**に過ぎないと考え，ユーザの訂正後に学習して語彙を増やしていく．このように，ユーザ自身も訂正作業により図 1(ii) の性能向上へ貢献することができる．
- さらに，これを個人的な訂正作業に留めずに，この**ユーザ参加**型の仕組みを発展させ，多数のユーザの訂正結果を Web サービス上で共有して性能改善を図る**社会的訂正**の枠組みを実現する．社会的訂正では，他の人々の利便性に貢献している実感が得られる上に，他のユーザが訂正している活動を見ることで，訂正の意欲が高まる可能性がある．これは**集合知** (wisdom of crowds) を利用して図 1(iii) のユーザ体験向上を実現するものである．

つまり音声認識研究 2.0 は，いわば**永久にベータ版** (perpetual beta) とも言える完全ではない音声認識に基づく Web サービスを，Web 上で多数のユーザの協力を仰ぎながら使ってもらうことで機能改善し，研究を進めていくアプローチとして位置付けられる．

我々は，このように図 1 を回していくことを目指し，音声認識研究 2.0 と Web 2.0 の両者の考え方に基づく Web サービス PodCastle (http://podcastle.jp) の一般公開を 2006 年 12 月 1 日から開始した [2)~4)]．

4 音声認識に基づくポッドキャスト検索サービス PodCastle

PodCastle[2)~4),11)~13)] は，ポッドキャストをテキストで検索，閲覧，編集できるソーシャルアノテーションシステムであり，同時に Web サービスの名称

☆1 もちろん我々自身も，音声認識研究 2.0 によって難易度の高い音声データに対する性能上の問題点をより一層自覚することで，音声認識研究 1.0 に継続的かつ積極的に取り組んでいく．

☆2 この点は，Web 2.0 にはない，音声認識研究 2.0 の大きな特長である．例えば，Wikipedia[10)] 等の集合知を利用した他の Web サービスでは，ユーザの貢献は編集した項目に限定され，自動的に他の項目へ波及して改善されることはない．

でもある．ポッドキャストには，一連のエピソードと呼ばれる音声データ（MP3 ファイル）に加え，その流通を促すために，ブログなどで更新情報を通知するために用いられているメタデータ RSS が必ず付与されている．エピソードは作成者（ポッドキャスタ）側で任意のタイミング（毎日，毎週等）で追加できる．この仕組みによりポッドキャストは音声版ブログとも言われ，個人による音声データの発信，流通，入手が容易にできる点が普及を促してきた．そして，Web 上のテキストに対して全文検索サービスが不可欠になったのと同様に，音声データに対しても PodCastle のような全文検索サービスの重要性が増している．

PodCastle よりも以前に，ポッドキャストを音声認識によりテキスト化し，ユーザが Web ブラウザ上で入力した検索語を含むポッドキャストの一覧を提示できる Web サービスとして，Podscope[14] と PodZinger[15] の二つが公開されていた．これらが英語の音声認識に基づくサービスであるのに対して，PodCastle は初めて日本語のポッドキャストに対する全文検索を実現するものであるが，言語の違いを除いても，以下の三つの点で本研究とは相違していた．

1. 従来は音声認識をしていても，表示される認識結果は一部に限定されており，音声を聞かずにポッドキャストの詳細な内容を把握できなかった．
2. 音声認識により索引付けされた全文テキストは内部に隠蔽され，外部のテキスト全文検索 Web サービスからは検索できなかった．
3. 音声認識にとって不可避な認識誤りが起きて検索に悪影響を与えていても，ユーザがそれらを訂正して改善することは不可能だった．

このように音声認識結果の完全開示による外部の検索サービスからの利用や，不特定多数のユーザの協力に基づく音声認識性能の向上を可能にするのは，我々の調査した範囲では PodCastle が初めてであった．

4.1 PodCastle の 3 つの機能

PodCastle では,「検索」,「閲覧」,「編集」の 3 つの機能を提供する Web サービスを一般公開しながら研究を進めることで，表 1 の音声認識研究 2.0 のすべての項目を満たし，図 1 のポジティブスパイラルを回していく．図 1 の (i) の性能理解は,「検索」機能と「閲覧」機能によって実現され，(ii) のユーザによる性能向上への貢献は,「編集」機能によって実現される．(iii) のより良いユーザ体験に結び付けるための性能向上については，訂正結果に基づく音響モデル，言語モデルの再学習等の様々な手法に取り組んでおり，紙面の制約から，詳細は文献 4),11)~13) に譲って省略する．以下，これら 3 つの機能の特長を述べる．

4.1.1 「検索」機能

音声認識結果，訂正結果の全文テキストを索引情報として使用して，全文検索する機能である．検索語

図 2: PodCastle の「編集」機能の画面表示例

をタイプすると，その語を含むエピソードの一覧が検索語付近のテキストと共に表示され，個々を試聴できる．そのうち一つを選択すると，次の「閲覧」機能に移行して全文テキストを読むことができる．

4.1.2 「閲覧」機能

検索したポッドキャストを「聞く」だけでなく，テキストで「読む」ことができる機能である．音声の再生に同期してテキスト中のカーソル（ハイライト）が動く．また，誤りを発見しやすいよう，音声認識時に推定した形態素ごとの信頼度に応じて着色される．

このように各エピソードの全文テキストは外部公開されているため，外部のテキスト全文検索サービスで，通常の Web ページと共に PodCastle のエピソード閲覧ページが発見される．その結果，ポッドキャストがより多くのユーザの目に触れて価値が高まる．これはポッドキャスタにとってもメリットがあるので，不特定多数のユーザに加え，ポッドキャスタ自身も次の「編集」機能で訂正する動機付けの一つとなる．

4.1.3 「編集」機能

ユーザが検索・閲覧中に認識誤りを発見したら，そのテキストを編集して「アノテーション」ができる機能である．各認識誤りの箇所において，競合候補の中から正しい候補を選択するか，正しいテキストをタイプして訂正する．そのために，閲覧時の全文表示画面とは別に，音声に同期してスクロールする図 2 のような画面で前後の見通し良く効率的な訂正ができる機能を用意した．これは，以前我々が提案した「音声訂正」[16] に基づくインタフェースであり，単語グラフを圧縮した confusion network（信頼度付き競合候補）を求めることで，候補表示を可能にした．

4.2 PodCastle の実装

PodCastle のシステム構成図を図 3 に示す．**Web クローラ**はポッドキャストを収集してデータベース管理部へ登録する．そして，認識処理を繰り返している複数の**音声認識器**から**音声認識状態管理部**へリクエストがあると，次に認識すべきエピソードが引き渡される．音声認識器がその認識処理を終えると，認識結果は音声認識状態管理部を経てデータベース管理部に渡される．**データベース管理部**では，ポッドキャ

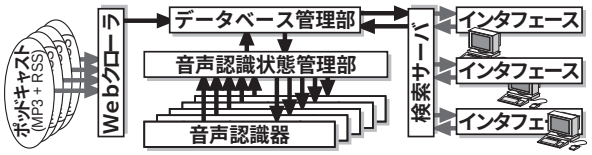

図 3: PodCastle のシステム構成図

ストとその音声認識結果，ユーザによる訂正情報を索引付けして，処理状態の管理をする．最後に，**検索サーバ**は，Web サイトとしての機能を持ち，ユーザによる検索と**インタフェース**の画面遷移を管理する．

5 議論

PodCastle は，Web の力を使って**集合知 (wisdom of crowds)** を利用するという Web 2.0 の原則 [1] を実践している．不特定多数のユーザによる訂正という参加のアーキテクチャを内在しており，ユーザの集合知によって，検索性能が改善していくポッドキャスト検索・閲覧（半自動書き起こし）サービスと捉えることができる．これらの改善がさらなるユーザによる貢献を促し，ユーザが増えるほど改善されるというソーシャルアノテーションのポジティブスパイラルが生まれる．各エピソードのすべての誤りを訂正する作業は労力が大きいが，そうした完全な訂正は求めずに，一部分だけでも訂正して貢献すると，性能が向上する仕組みになっていることが重要となる．

我々は Web 2.0 の「ユーザを信頼する」立場から，基本的にはユーザによる訂正の質は高いものと考えており，実際に公開後に集まった訂正結果の質は高い．しかし，もし仮にユーザが故意に不適切な訂正をした場合のために，その信頼性を音響尤度に基づいて評価する方法も検討している．また，利便性の向上も兼ねてユーザ識別機能 (ID) を導入する予定である．

今後，マッシュアップ用の各種 API を整備すると共に，この研究アプローチを幅広く展開するために，我々以外の研究者も研究開発に参加できる枠組みを検討中である．それ以外にも，音声認識性能の改善等，様々な拡張の余地がある．例えば，語学学習用ポッドキャスト等で，言語識別機能や他言語への対応が必要なことが判明している．語句の読みをユーザが明示的に教えられるインタフェースや，動画中の音声にも対応予定である．また，個人が持つインタビューや会議等の音声を書き起こすために，PodCastle のインタフェースを活用できるようにすることで，訂正作業に強い動機を持つユーザの参加も促せるようにしていきたい．

6 おわりに

これまでの音声認識研究と相補関係にある「音声認識研究 2.0」という新たな研究アプローチと，その実例として，集合知を活用した音声情報検索用 Web サービス「PodCastle」を紹介した．本研究の学術的意義は，不特定多数のエンドユーザに音声認識誤りを訂正する協力をしてもらうことで，音声認識・音声情報検索の性能をどこまで高くできるかを探求することにある．同時に，日本語ポッドキャスト検索のための世界初の Web サービスを公開して，エンドユーザの役に立つという社会的意義も持っている．

さらに本研究は，音声コーパスの用意が困難な状況で，どのようにすれば音声認識が役に立つかを明らかにする点でも意義がある．コーパスの整備は多大なコストと労力を要する上に，適用範囲が限定される問題があったが，本研究では，誤認識も含めて全テキストを外部公開し，ユーザの訂正によって「音声認識を育ててもらう」方針を取った．この場合，誤認識が多いために批判を受けるリスクはあるが，そうした現状をユーザと共有してはじめて，音声認識技術の真の普及と発展があると我々は考える．本研究により，ユーザの貢献を積極的に取り込んで音声認識の実用化へ向けて研究する重要性と将来性が明らかになり，多くの研究者が取り組むことで，今後の音声認識・音声情報検索の研究分野に新たな展開を引き起こすことができればと願っている☆3.

謝辞 Web サーバとクライアントの実装を担当して頂いた有限会社ブラジル（上津 竜太郎 氏），有限会社メロートーン（新井 俊一 氏），沢田 洋平 氏に感謝する．

参 考 文 献

[1] O'Reilly: What Is Web 2.0 — Design Patterns and Business Models for the Next Generation of Software, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.
[2] 後藤, 緒方, 江渡: PodCastle の提案: 音声認識研究 2.0 を目指して, 情処研報音声言語情報処理 2007-SLP-65-7 (2007).
[3] Goto, Ogata, and Eto: PodCastle: A Web 2.0 Approach to Speech Recognition Research, *Proc. of Interspeech 2007* (2007).
[4] 緒方, 後藤, 江渡: PodCastle: ポッドキャストをテキストで検索，閲覧，編集できるソーシャルアノテーションシステム, WISS 2006 論文集 (2006).
[5] 嵯峨山: なぜ音声認識は使われないか・どうすれば使われるか？, 情処研報音声言語情報処理 94-SLP-1-4 (1994).
[6] 中川: 音声言語処理の進歩と今後, 情処研報 音声言語情報処理 2004-SLP-50-4 (2004).
[7] 畑岡: 音声技術実用化の課題と取り組み, 情処研報 音声言語情報処理 2005-SLP-55-1 (2005).
[8] 赤堀ほか: パネルディスカッション「音声認識技術の実用化」, 情処研報音声言語情報処理 2005-SLP-58-6 (2005).
[9] 石川ほか: パネルディスカッション「音声認識の実用化の阻害要因と課題」, 情処研報音声言語情報処理 2006-SLP-63-9 (2006).
[10] Wikipedia: http://www.wikipedia.org/.
[11] 緒方, 後藤, 江渡: PodCastle の実現: Web 2.0 に基づく音声認識性能の向上について, 情処研報 音声言語情報処理 2007-SLP-65-8 (2007).
[12] Ogata, Goto, and Eto: Automatic Transcription for a Web 2.0 Service to Search Podcasts, *Proc. of Interspeech 2007* (2007).
[13] 緒方, 後藤, 江渡: PodCastle: Web 2.0 に基づくポッドキャスト音声認識手法, 音講論集 秋季 1-3-5 (2007).
[14] Podscope: http://www.podscope.com/.
[15] PodZinger: http://www.podzinger.com/.
[16] 緒方, 後藤: 音声訂正：選択操作による効率的な誤り訂正が可能な音声入力インタフェース, 情処学論, **48**, 1, (2007).
[17] 西村ほか: ネットワーク公開試験に向けた音声対話 Web アプリケーションの開発, 音講論集 春季 1-9-9 (2007).

☆3 2007 年 3 月からは，本研究とは独立に，w3voice Laboratory[17] の公開試験も始まっており，今後も多くの研究者が音声認識に基づく Web サービスを公開していくことを期待したい．