特集論文

PodCastle: ユーザ貢献により性能が向上する音声情報検索システム

PodCastle: A Spoken Document Retrieval System Improved by User Contributions

後藤 真孝 Masataka Goto 産業技術総合研究所 National Institute of Advanced Industrial Science and Technology (AIST) m.goto@aist.go.jp, http://staff.aist.go.jp/m.goto/

緒方 淳 Jun Ogata （同 上）

江渡 浩一郎 Kouichirou Eto （同 上）

keywords: information retrieval, speech recognition, error correction, wisdom of crowds, Web 2.0

Summary

In this paper, we describe a public web service, *"PodCastle"*, that provides full-text searching of speech data (Japanese podcasts) on the basis of automatic speech recognition technologies. This is an instance of our research approach, *"Speech Recognition Research 2.0"*, which is aimed at providing users with a web service based on Web 2.0 so that they can experience state-of-the-art speech recognition performance, and at promoting speech recognition technologies in cooperation with anonymous users. PodCastle enables users to find podcasts that include a search term, read full texts of their recognition results, and easily correct recognition errors by simply selecting from a list of candidates. Even if a state-of-the-art speech recognizer is used to recognize podcasts on the web, a number of errors will naturally occur. PodCastle therefore encourages users to cooperate by correcting these errors so that those podcasts can be searched more reliably. Furthermore, using the resulting corrections to train the speech recognizer, it implements a mechanism whereby the speech recognition performance is gradually improved. Our experience with this web service showed that user contributions we collected actually improved the performance of PodCastle.

1. はじめに

インターネット上のデータ量が増大し，情報検索は社会にとって不可欠な技術となった．既にテキスト（文字）データの検索に関しては，多数の Web サービスが公開されて普及し，キーワードをタイプするだけで全文検索が可能となっている．音声データに関しても，近年，音声版のブログ (Weblog) ともいえる「ポッドキャスト」や音声を伴う動画等がインターネット上で急増しており，音声を含むデータに対する情報検索の重要性が増しつつある．しかし，テキストデータと異なり，音声データ自体を索引として使えないため，その実現は難しい．テキスト同様に全文検索サービスを可能にするには，音声認識によるテキスト化（書き起こし）が必要だが，従来の音声認識技術では，誤認識が多い上に，新しい言葉に対応できなかったため，インターネット上の多様な音声に対して，実用的な精度の音声情報検索サービスはできなかった．

そこで本研究では，Web 上の日本語のポッドキャストを音声認識によって自動的にテキスト化することで，それらをユーザが全文検索できるだけでなく，詳細な閲覧，編集も可能な音声情報検索システム**「PodCastle」（ポッドキャッスル）**を提案する．PodCastle では，検索したポッドキャストの全文をテキスト表示することで，音声再生環境がなければ内容を把握できないポッドキャストを「読む」ことも可能にする．従来こうしたシステムが実現困難だったのは，ポッドキャストの多様な音声に対して，高い音声認識率を達成することが難しかったからである．本研究では，これを解決するために，すべての音声認識結果を積極的にユーザに開示し，多数のユーザに認識誤りを訂正（認識結果を編集）する協力をしてもらうことで，音声検索性能と音声認識性能をシステムの運用中に向上させる枠組みを提案する．これにより一般ユーザだけでなく，作成者（ポッドキャスタ）が自身のポッドキャストの認識誤りを把握し，訂正して適切に検索されるように改善できる．

そうしたユーザの自発的な協力を得るために，PodCas-

tle では音声認識結果のテキスト全文を表示するだけでなく，認識誤りを容易に訂正できる編集（アノテーション）機能を提供する．閲覧中に認識誤りをユーザが発見してマウスやキーボードで指摘すると，その競合候補（音声認識中に，最終的な認識結果以外に可能性の高かった単語候補）が自動的に表示される．ユーザはその周辺の音声を聞きながら，正しい候補を選択するか，正しいテキストをタイプすることで訂正する．これにより単にテキストエディタで訂正するのに比べ，効率が良く，気軽に作業ができる．以上のシステムを Web 上で公開して，多数のユーザが検索，閲覧，編集を繰り返すと，誤認識箇所に関する正解情報が日々集積されるので，それを自動学習することで音声認識性能も向上できる．つまり，個々の訂正の影響はそのポッドキャストにとどまらず，未訂正のポッドキャストに対する性能向上も可能となる．こうした多数のユーザによる「ソーシャルアノテーション」を利用したアプローチは，従来の音声認識研究では実現されていなかった新たな性能向上の枠組みと言える．

このように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 Web 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我々は**「音声認識研究 2.0」**と名付ける．本研究は,「PodCastle」という音声情報検索システムの実現自体が重要で一つの貢献となっているだけでなく,「音声認識研究 2.0」という Web 技術を音声認識研究で活用するための新しい研究アプローチを提案する点も貢献となっている．そこで，以下ではまず，2 章 で音声認識研究 2.0 として提案する研究アプローチについて議論し，具体的に Web 技術を活用する方法を示す．次に 3 章で，その実例として Web サービス PodCastle を提案する．そして，4 章で Web 2.0 との関連を考察し，今後の展望を議論する．最後に，5 章で本研究の意義を総括する．

2. 音声認識研究 2.0

「音声認識研究 2.0」とは，不特定多数のエンドユーザの協力を仰ぎながら，音声認識の性能向上と実用化（利用率の向上）を共に実現することを目指した，音声認識の新たな研究アプローチである．以下，その特長を議論した上で，どのように可能にしていくかを述べる．

2·1 二つの特長

音声認識研究 2.0 は，(1) ユーザに対して音声認識の現状を積極的に開示し，(2) ユーザの協力を得て音声認識技術を発展させていく，という二つの特長を持つ．

一般に，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，そのときの印象で音声認識の有効性に疑問を抱き，使わなくなることが多い．様々な研究により音声認識率が向上し，文献 [嵯峨山 94] 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 [中川 04]，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 [畑岡 05, 赤堀 05, 石川 06].

音声認識研究 2.0 では，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すことができるという一つ目の特長を持つ．その実例として，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 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．

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．

音声認識研究 2.0 では，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という二つ目の特長を持つ．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．

2·2 ポジティブスパイラル

音声認識研究 2.0 で性能向上と利用率向上を共に実現するには，図 1 のポジティブスパイラルを回す，つまり三段階が繰り返されることが重要である．従来は，音声認識の普及のために重要なこの三段階のそれぞれに阻害要因があったため，これが回っていなかったと考えられる．

- (i) の性能理解に関しては，従来は，ユーザ自身の発

(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）

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，ユーザとしての自分の発声ではなく，他人の適切な発声（コーパス中の音声）を認識した結果を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[*1]．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．例えば，研究目的で音声認識ソフトウェア（例えば文献 [李 05]）が公開されても，主に開発者向けでエンドユーザが直接利用する機会は少なく，音声対話システム（例えば文献 [鹿野 06]）が街中に設置されても，その地域を訪れたユーザしか体験できなかった．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，これらの問題を解決することで，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．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る．これはあくまで,「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．これは Web 2.0 [O'Reilly] を意識して付けた名称であり，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．

[*1] ただし，ユーザに意識させることなく利用中に自動的に話者適応したり，研究レベルでは未知語を自動獲得したりできるシステムも存在する．しかし，いずれの場合も，それらがエンドユーザ間で共有されることはなかった．

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 [O'Reilly] 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．

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，ポッドキャスト等の **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る．また，認識用の辞書に登録されていない未知語も多くなる．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．未知語に関しても，ユーザがまだアノテーション（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，専門家である研究者だけでなく，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) の試験公開を 2006 年 12 月 1 日から開始した [緒方 06].

3. 音声認識に基づくポッドキャスト検索サービス PodCastle

PodCastle 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 (Really Simple Syndication) が付与されている．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 [Podscope] と PodZinger [PodZinger] の二つが公開されていた．Podscope では，ポッドキャストのタイトルだけが列挙され，音声認識結果のテキストは一切表示されないものの，検索語が出現する箇所は再生できる．一方，PodZinger では，これに加え，検索語が出現した周辺のテキストも表示され，ユーザが内容を把握しやすくなっている．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．

3·1 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によって実現される．以下，これら 3 つの機能を説明する．

§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．画面表示例を図 2 に示す．一般的なテキスト全文検索サービスのように検索語をタイプすると，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，エピソードごとに用意された「再生」「停止」ボタンを押して試聴できる．検索語の出現箇所は着色されており，そのうち一つを選択すると，次の「閲覧」機能に移行して，全文テキストを選択した検索語付近から見ることができる．

§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．画面表示例を図 3 に示す．表示では，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．再生速度を，10 %単位で再生中に自在に早くしたり遅くしたりできる機能も

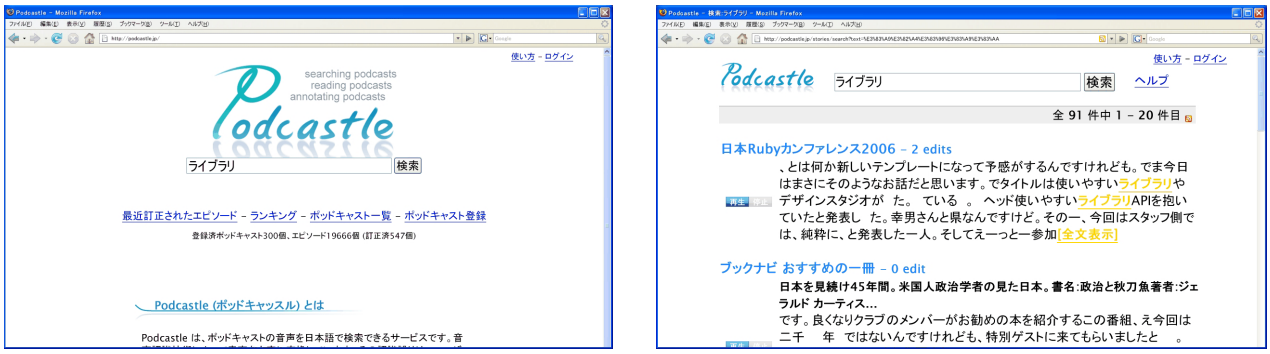

図 2 PodCastle の「検索」機能の画面表示例: 左のトップページの画面でキーワードをタイプ入力すると，右のような全文検索結果の画面が表示される．

図 3 PodCastle の「閲覧」「編集」機能の画面表示例: 全文検索結果から興味のあるポッドキャストを選択すると，左の詳細表示画面になり，検索したキーワード周辺の音声認識結果を再生しながら見ることができる．区間ごとに一番上が認識結果，その下に並んでいるのが可能性の高い候補であり，適切な候補を選択するだけで訂正できる．右の図のような全文表示画面にも切り替えられる．全文表示画面では，音声認識の信頼性が低い区間が赤色で着色されており，誤り箇所（単語）をクリックするとその下に選択用の候補が表示される．詳細表示画面と全文表示画面は，訂正中のカーソル位置がそのままの状態で，再生中でも自由に切り替えられる．

用意した．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赤色で着色され，訂正済の区間は水色で着色される．

ポッドキャストの音声再生は魅力的である一方，音声であるために，その内容に関心があるかどうかを聞く前に把握することは容易でなかった．再生スピードを上げることで聞く時間を短縮しようとしても限界がある．本機能により，聞く前にざっと全文テキストを眺められることで，内容に対する関心を短時間で判断でき，ポッドキャストの取捨選択が効率良くできる．また，収録時間の長いポッドキャストでは，どの辺に関心のある部分があるのかを見つけて，そこから聞くことができて便利である．仮に音声認識誤りが含まれていても，こうした関心の有無は判断できることが多い．

本機能により，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（ポッドキャスト作成者）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．

§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．ここでのアノテーションは，ポッドキャストに対して書き起こしテキストを作成することを意味し，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，閲覧に適した図 3 右側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 3 左側の詳細表示画面で，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，文献 [緒方 07a] の「音声訂正」に基づくインタフェースであり，音声認識の内部状態である単語グラフ（認識結果の膨大な可能性の探索結果）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．この表示で，競合候補の個数が多い箇所は認識時の曖昧性が高かった（音声認識時に自信がなかった）ことを表しており，候補の個数に注意しながら作業することで，誤り箇所を見逃しにくい．各区間の競合候補は信頼度の高い順に並んでおり，通常は上から下へ候補を見ていくと，早く正解にたどり着けることが多い．また，競合候補には必ず空白の

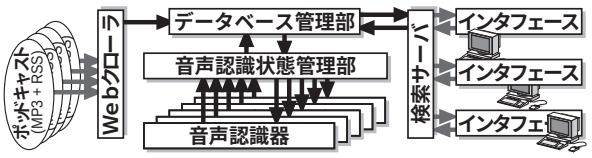

図 4 PodCastle のシステム構成図

候補が含まれる．これは「スキップ候補」と呼ばれ，その区間の認識結果をないものとする役割を持つ．つまりこれをクリックするだけで，余分な単語が挿入されている箇所を容易に削除できる．各訂正操作は取り消し（アンドゥ）ができ，保存ボタンを押したときだけ，そのエピソードに関するすべての訂正結果がアップロードされて，PodCastle のデータベースに蓄積される．

3·2 PodCastle の実装

PodCastle のシステム構成図を図 4 に示す．**Web クローラ**は RSS が登録されている各ポッドキャストにおいて，更新されたエピソード（音声データの MP3 ファイル）を定期的に収集して，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，ポッドキャストとその音声認識結果 (confusion network)，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．なお，ユーザが PodCastle を利用中に再生される音声データ（MP3 ファイル）は，PodCastle を経由せずに元のポッドキャスト配信サイトから，ユーザのブラウザ（クライアント）へ直接ダウンロードされる．

音声認識器には，confusion network（信頼度付き競合候補）を生成できる特殊な機能を持つ大語彙連続音声認識器 [緒方 07d, Ogata 07b, 緒方 07c] を用いた．音声認識の内部状態を適切な個数の競合候補にまとめ上げるのは難しく，通常の音声認識器では候補表示ができないためである．本認識器は back-off 制約 N-best 探索アルゴリズム [緒方 01] に基づいており，その結果の単語グラフに対して，consensus デコーディング [Mangu 00] を行い，confusion network を生成する．ポッドキャストには純粋な音声のみのデータ以外にも，騒音下での音声データや，背景に音楽が重畳している音声データなども多く存在する．そこで，音声認識の前処理として，GMM を用いた音声，音楽，無音の 3 種類の音響イベント検出によって，認識すべき音声発話区間を推定し，雑音対処のために ETSI Advanced Front-End [ETSI 02] を音響分析に適用して性能を改善した．

音声認識器の言語モデルとしては，できるだけ多くの語彙が認識対象となるように，毎日新聞記事 10 年分（1991 年〜2001 年）のテキストデータと，日本語話し言葉コーパスの 2670 講演分の書き起こしデータを，基本的な学習用コーパスとして利用した．しかしポッドキャストの場合，最新の話題や新しい言葉（新語，時事用語，芸能人名，固有名詞等）を含むものが多く，従来の音声認識技術では対応できない．従来の技術では，事前に用意した音声認識辞書の語句しか認識できないため，辞書にない新しい言葉は，既存の何らかの語句の組み合わせとして誤認識されてしまうからである．また，手作業で言葉を登録するのは対象が膨大すぎて現実的でないだけでなく，その前後の文脈が得られないために性能が低下することがある．そこで我々は，インターネット上のニュース記事や辞書から，そうした新しい言葉を日々収集して自動学習する新たな技術を開発した．具体的には，Yahoo! ニュースに掲載された記事と，不特定多数のユーザの協力で構築されているインターネット辞書「はてなキーワード」[はてな] を用いて，新しい言葉を自動学習する仕組みを実現した．その際，音声認識辞書に言葉を追加するだけでなく，その前後の文脈（n-gram 確率）も学習し，より的確な認識が可能となった．

一連の機能のサーバ側動作は，Web アプリケーションフレームワーク Ruby on Rails で実装されている．プログラミング言語 Ruby，Web サーバ Mongrel，データベース MySQL を用い，ポッドキャストのメタデータと音声認識結果をデータベースに登録して運用した．confusion network の各候補をパースし，データベースにレコードとして登録する．ユーザがブラウザ（クライアント）上で検索語を入力すると，サーバ側で Senna (Tritonn) の全文検索機能によって検索し，検索結果のページを表示する．クライアント側のインタフェース（画面描画，カーソル移動，候補表示等）の機能は，JavaScript 及び JavaScript ライブラリ MochiKit を用いて実装した．ただし，音声ファイル中の任意の位置からの再生や再生速度の変更は，JavaScript だけでは実現できないため，Quicktime プラグイン，Flash プラグインと連携させることによって実現した．なお，ユーザの利便性向上のためにユーザ識別（ユーザ ID）機能も既に実装しており，既存の Web 認証 API として OpenID に対応している．匿名状態でも使用できるが，ログインすることにより，過去の自分の訂正履歴を閲覧でき，過去に途中まで訂正したエピソードの続きを作業することが容易となっている．また，「訂正回数の多いユーザ」のランキングにも表示される．

3·3 PodCastle の運用結果

PodCastle (http://podcastle.jp) の試験公開を 2006 年 12 月 1 日から開始し [緒方 06]，2008 年 6 月 12 日にプレス発表をして一般公開による実証実験を開始した．2009 年 7 月 29 日時点で，登録済のポッドキャスト数は 571 件，エピソード (MP3) 数は 52917 件，そのうち一部でも訂正

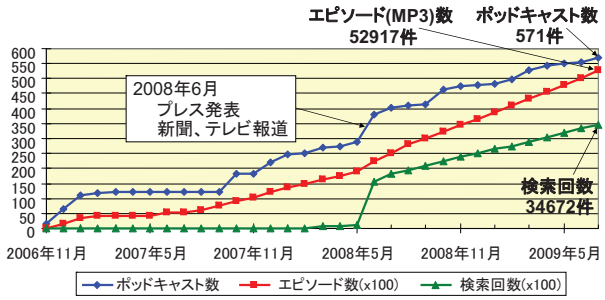

図 5 PodCastle の利用状況: 登録済のポッドキャスト数，エピソード (MP3) 数，検索回数の累積件数

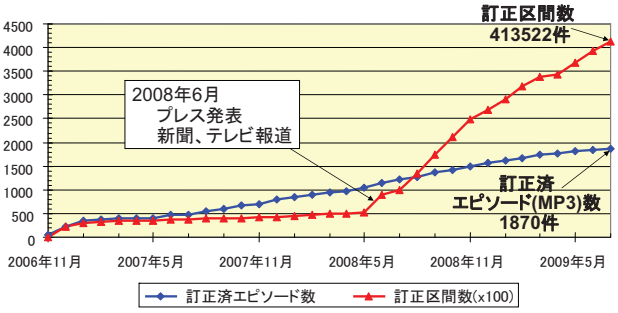

図 6 PodCastle の利用状況: 訂正済エピソード数，訂正区間数の累積件数

されたエピソード数は 1870 件（全エピソードの 3.53 %）であった．ポッドキャストの登録件数の増加の様子を図 5 に示す．2008 年 6 月に多数の報道がなされた結果，利用がさらに延びていたことがわかる．また，訂正済エピソード数の増加の様子を図 6 に示す．訂正済エピソード数は，一部でも訂正がなされれば集計されるが，一つのエピソード内でどの程度訂正されているかはわからない．そこで，実際に訂正がなされた区間数の総計 (413522 件) も示した．基本的にユーザは匿名で訂正しており，これが何人のユーザによる訂正かは不明である．しかし，3·2 節で述べたようにユーザは希望すれば OpenID によってログインすることができ，「訂正回数の多いユーザ」のランキングを表示する機能があるので，その上位 10 ユーザによる訂正区間数を求めたところ，42965 件であった．これは訂正区間数の総計 (413522 件) の 10.39 %に相当するため，それ以外のユーザによる訂正の貢献は 89.61 %とみなすことができる*2．

図 6 の訂正区間数から，特に 2008 年 6 月の報道以降，ユーザによる訂正が増えていることがわかる．実際に，複数のポッドキャストにおいて，新規エピソードが配信されると短期間で，ほぼすべての認識誤りが訂正される現象が観測されており，週一回配信される芸能人や声優によるポッドキャストだけでなく，毎日配信されるニュースのポッドキャストまで頻繁に訂正されている現象が起きていた．

*2 ただし，上位 10 ユーザがログインをしない状態で訂正していた可能性もあるため，あくまで目安に過ぎない．

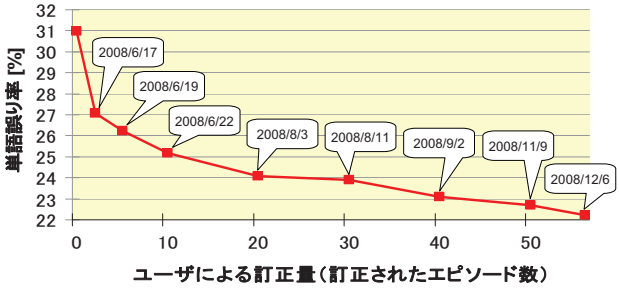

図 7 学習の効果の例: ユーザによる訂正量（あるポッドキャストに対する訂正されたエピソードの総数）の増加に伴う単語誤り率の減少（吹き出し部分は，そのエピソード数の訂正が得られた日付を示す）

実際に，その訂正を学習に利用することで，音声認識の性能を向上させることが可能であることを実験的に確認している [Ogata 09a, 緒方 09b]．そうした性能向上の具体例として，図 7 に，あるポッドキャストにおける学習の効果を示す．この図から，ユーザによる訂正量 (訂正されたエピソードの総数) が増加するにつれて，単語誤り率が減少して性能が向上していることがわかる．本ポッドキャストでは週一回エピソードが配信されており，それに追従して，ほぼすべてのエピソードに対して多くの訂正がなされていた．なお上記は，音声認識の音響モデルに関して学習した結果であり，具体的な学習方法の詳細やその他の実験結果については，文献 [Ogata 09a, 緒方 09b] を参照されたい．また，言語モデルの学習効果の評価に関しては今後の課題であり，稿を改めて報告予定である．このように，本ポッドキャストの事例に限らず，訂正が多くなされる前は頻繁に誤っていたポッドキャストでも，訂正量が増えるにつれて，新しく配信されるエピソードの認識結果ではあきらかに認識誤りが減り，ユーザの訂正の負担や訂正時間が減少していたことがわかった．

なぜ訂正したくなるのかという動機については，不特定多数のユーザが訂正している関係上，直接調査するのは困難だが，我々が考察したりインターネット等で情報収集したりした結果から，主に以下の四つの理由があると推測される．

- 面白いから

  PodCastle が提供する訂正インタフェースは，操作が快適なように注意深く実装されており，訂正操作自体が面白いものとなっている．実際に慣れてくると，誤りを発見してすばやく訂正する操作を，ゲーム感覚で楽しむことができる．
- 貢献したいから

  自分の利便性を向上させたいという動機だけでなく，音声認識，音声検索の性能向上に貢献できること自体が嬉しいという気持ちから，自発的に訂正することがある．
- 検索されて欲しいから

  主に個人のポッドキャストの作成者（ポッドキャ

スタ）が，自分の配信しているポッドキャストの誤認識を減らし，より的確に検索・閲覧されるように訂正することがある．

- 誤認識を許せないから

  自分が好きなポッドキャストの認識結果に，誤認識がそのまま残っているのが許せず，誤認識を訂正することがある．主に芸能人のポッドキャスト等で，そのファンがこうした動機で訂正することがあると考えられる．

4. 議　　論

音声認識研究 2.0 は，音声認識に Web 2.0 の考え方を導入したものであるが，以下では，その実例である PodCastle がどの点で Web 2.0 的と言えるのかを考察し，様々な観点から今後の展望を議論する．

4·1 PodCastle と Web 2.0 との関連

PodCastle という Web サービスと，Web 2.0 との関連性を考察する．Web 2.0 [O'Reilly] は，Tim O'Reilly らによって 2004 年に提唱された概念で，近年，Web 上の一連の新しい潮流を包含して説明する際に用いられることが多い．そこで以下では，インターネットアプリケーションである PodCastle が持つ特長の中で，Web 2.0 の考え方に基づいているものを列挙する．

- **集合知 (wisdom of crowds)，参加のアーキテクチャ，ユーザによる貢献**

  PodCastle は，Web の力を使って集合知を利用するという Web 2.0 の原則を実践している．不特定多数のユーザによる誤認識箇所の訂正が前提であり，そうしたユーザの参加を促すアーキテクチャを内在している．PodCastle は，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サービス，かつ，認識性能が改善していくポッドキャスト閲覧（半自動書き起こし）サービスと捉えることができる．そして，これらの改善がさらなるユーザの参加を促し，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．

  ただし，ここで重要なのは「参加」つまり「訂正」の仕方と質である．音声認識性能によっては，ポッドキャストのエピソードを最初から最後まで訂正する作業は労力が大きく，数時間かかることがある．そこで PodCastle では，そうした完全な訂正は求めずに，ユーザの気付いた範囲，可能な範囲の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．つまり，少数の人から多大な貢献を期待するのではなく，多数の人から少しずつの貢献を期待する立場を取る．一方，訂正の質の問題については，4·2 節 で改めて議論する．

  ユーザによる貢献に関してさらに議論を深めると，本研究で「ユーザの貢献を増幅」する枠組みを実現した点が，通常の Web 2.0 にはない「PodCastle ならでは」の大きな特長となっている．例えば，Wikipedia [Wikipedia] 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．それに対して PodCastle では，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という技術を初めて実現した．この「ユーザの貢献を増幅して性能向上へ繋げる技術」により，ユーザが貢献（訂正）していない箇所へ波及して改善される点が重要である．

- **ロングテール (long tail)**

  有名なポッドキャストと有名でなく通常は発見されにくいポッドキャストは対等なので，PodCastle 上の検索結果として表示されることで聴取が促される．さらに，3 章でも述べたように，全文テキストを公開することで，Google 等の外部の一般的なテキスト検索エンジンから検索されることも意図している．PodCastle のサイト上で，ユーザは新たなポッドキャストの RSS の URL を自由に登録できるので，さらに検索対象が拡大し，豊かなテールを築いていける．

- **パーマリンク (permalink)**

  PodCastle では，ポッドキャストやそれを構成する各エピソードの URL は，パーマリンクとしてユーザが外部利用できることを重視している．これによりある特定のポッドキャストについて言及したいときに，RSS や MP3 ファイルの URL でなく，その全文テキストが見られる PodCastle のパーマリンクを利用できて便利である．

- **RSS の配信**

  PodCastle 上で特定の検索語を含むエピソードを検索できるだけでなく，その検索語を含むエピソード群を購読し続けるための RSS を配信する機能も既に提供している．ただし，RSS には様々なフォーマットがあり，すべてに対応することは難しい．そこで，PodCastle では最小限の RSS を配信し，あとはユーザ側でマッシュアップしてもらうことを期待することとする．例えば，Plagger [Plagger] 等の外部のカスタマイズ可能なフィードアグリゲータで，最小限の RSS を利用して，任意の形式の RSS や付加情報を持つ RSS を生成して使うことができる．

4·2 今 後 の 展 望

PodCastle は，Web 2.0 の「永久にベータ版」という考え方に基づき公開を開始したため，今後も以下に述べるような様々なアイディアで拡張を続けていく予定である．

- **マッシュアップ (mashup)，フォークソノミー (folksonomy)**

Web 2.0 が持つ重要な概念の中でまだ対応していないものに，マッシュアップとフォークソノミーがある．マッシュアップ用の各種 API に関しては，PodCastle 側で整備することを検討中である．フォークソノミーに関しては，各ポッドキャスト，エピソードに対するタギング（ユーザによる任意のキーワードでのラベル付け）への対応も検討したが，各エピソードはパーマリンク化してあるため，タギングをサポートした他のソーシャルブックマーク用 Web サービス等とマッシュアップした方が，ユーザの利便性が高いと考えている．

- **ユーザによる訂正の質（いたずら対策）**

  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（いたずら）をした場合には問題になるため，その信頼性を音響的に評価する方法の研究も進めている．例えば，訂正結果の中で読みが判明する箇所に関して音響信号とのアラインメントを求め，その音響尤度が低すぎたら信頼性の低い訂正結果と判定する方法等を検討している．

- **個人的な書き起こし用インタフェースへの対応**

  インタビューや会議，講演等を書き起こす需要は高く，PodCastle はそのために有用だが，それらの音声データはポッドキャストとして公開できないことが多い．そこで，そうした訂正作業に強い動機を持つユーザの参加を促すために，他のユーザには開示されないアクセス制限をかけるオプションを用意することを検討している．ここで重要なのは，その場合でも，訂正結果は全体の性能向上に寄与し，逆にその恩恵も受けられることである．

上記以外にも，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．

5. お わ り に

本論文では，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．一般に，充分なコーパスが用意できれば音声認識技術は有用であるが，そ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．それに対して本研究では，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．

謝　　辞

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

◇ 参 考 文 献 ◇

[赤堀 05] 赤堀 一郎, 渡辺 隆夫, 河井 恒, 庄境 誠, 畑岡 信夫：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, pp. 31–40 (2005)

[ETSI 02] ETSI standard document: ETSI ES 202 050 v1.1.1: Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms (2002)

[畑岡 05] 畑岡 信夫：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, pp. 1–6 (2005)

[石川 06] 石川 泰, 神沼 充伸, 中川 聖一, 磯 健一, 新田 恒雄：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, pp. 45–54 (2006)

[李 05] 李 晃伸：大語彙連続音声認識エンジン Julius の開発の進展, 情処研報音声言語情報処理 2005-SLP-59-22, pp. 127–132 (2005)

[Mangu 00] Mangu, L., Brill, E., and Stolcke, A.: Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks, *Computer Speech and Language*, Vol. 14, No. 4, pp. 373–400 (2000)

[中川 04] 中川 聖一：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, pp. 23–30 (2004)

[緒方 01] 緒方 淳, 有木 康雄：大語彙連続音声認識における最ゆう単語 back-off 接続を用いた効率的な N-best 探索法, 信学論 (D-II), Vol. J84-D-II, No. 12, pp. 2489–2500 (2001)

[緒方 06] 緒方 淳, 後藤 真孝, 江渡 浩一郎：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, pp. 53–58 (2006)

[緒方 07a] 緒方 淳, 後藤 真孝：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, Vol. 48, No. 1, pp. 375–385 (2007)

[Ogata 07b] Ogata, J., Goto, M., and Eto, K.: Automatic Transcription for a Web 2.0 Service to Search Podcasts, in *Proc. of Interspeech 2007* (2007)

[緒方 07c] 緒方 淳, 後藤 真孝, 江渡 浩一郎：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)

[緒方 07d] 緒方 淳, 後藤 真孝, 江渡 浩一郎：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, pp. 41–46 (2007)

[Ogata 09a] Ogata, J. and Goto, M.: PodCastle: Collaborative Training of Acoustic Models on the Basis of Wisdom of Crowds for Pod-

cast Transcription, in *Proc. of Interspeech 2009* (2009)
[緒方 09b] 緒方 淳, 後藤 真孝：PodCastle: ポッドキャスト音声認識のための集合知を活用した音響モデル学習, 第 3 回音声ドキュメント処理ワークショップ講演論文集, pp. 91–96 (2009)
[O'Reilly] O'Reilly, T.: 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
[Plagger] Plagger: http://plagger.org/
[Podscope] Podscope: http://www.podscope.com/
[PodZinger] PodZinger: http://www.podzinger.com/
[嵯峨山 94] 嵯峨山 茂樹：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, pp. 23–30 (1994)
[鹿野 06] 鹿野 清宏, Tobias, C., 川波 弘道, 西村 竜一, 李 晃伸：音声情報案内システム「たけまるくん」および「キタちゃん」の開発, 情処研報 音声言語情報処理 2006-SLP-63-7, pp. 33–38 (2006)
[Wikipedia] Wikipedia: http://www.wikipedia.org/
[はてな] はてなキーワード：http://d.hatena.ne.jp/keyword/

〔担当委員：高間 康史〕

2009 年 5 月 6 日 受理

著 者 紹 介

後藤 真孝

1998 年早稲田大学大学院理工学研究科博士後期課程修了. 博士（工学）. 同年, 電子技術総合研究所に入所し, 2001 年に改組された産業技術総合研究所において, 現在, 情報技術研究部門メディアインタラクション研究グループ長. 筑波大学大学院 准教授（連携大学院）, 統計数理研究所客員教授, IPA 未踏 IT 人材発掘・育成事業 未踏ユースプロジェクトマネージャーを兼任. ドコモ・モバイル・サイエンス賞 基礎科学部門 優秀賞, 科学技術分野の文部科学大臣表彰 若手科学者賞, 情報処理学会 長尾真記念特別賞等, 24 件受賞.

緒方 淳

2003 年龍谷大学理工学研究科博士後期課程修了. 同年, 産業技術総合研究所に入所し, 現在に至る. 博士 (工学). 音声認識, 音声インタフェースに関する研究に従事. 2000 年日本音響学会粟屋潔学術奨励賞, 2001 年電子情報通信学会学術奨励賞, 2004 年 WISS2004 ベストペーパー賞, 2006 年 WISS2006 ベストペーパー賞, 2006 年情報処理学会山下記念研究賞各受賞. 電子情報通信学会, 情報処理学会, 日本音響学会各会員.

江渡 浩一郎

1997 年慶應義塾大学大学院政策・メディア研究科修了. 修士（政策・メディア）. 同年, 国際メディア研究財団に所属. メディア・アーティストとして作品制作を行う. sensorium プロジェクトでアルスエレクトロニカ賞グランプリ, 同アルスエレクトロニカ賞 Honorary Mention を受賞. 2002 年独立行政法人産業技術総合研究所に所属. 現在, サービス工学研究センター最適化研究チーム研究員. 主な著書に「パターン, Wiki, XP」（技術評論社）. 日本ソフトウェア科学会, 情報処理学会, 各会員.