日本音響学会誌 68 巻 10 号 (2012), pp. 527–532

解説

Webで生きる/活きる音声認識*

緒方 淳, 後藤真孝 (産業技術総合研究所)**

43.72.Ne

1. はじめに

インターネット (Web) は, 膨大で多様な情報が日々集積する, 社会的に必要不可欠な知の基盤であるといえる。統計的手法が主流である現在の音声研究では, 膨大な情報・データを処理することが有効であることから, 研究開発における Web の利活用が注目されている。また, 近年, クラウドコンピューティングが進展し, Web を通じて計算サーバ, ストレージといった大規模な IT リソースを柔軟に利用できるようになったことで, そうした膨大なデータの処理・解析を行うための計算基盤も整ってきている。一方, Web は膨大なデータの集積場であるだけでなく, エンドユーザと技術の関わり (利用形態) に変革をもたらすものでもある。従って, Web は音声研究における一つの重要なプラットフォームになりつつある。

我々はそうした Web を利活用する音声認識研究プロジェクト「**PodCastle**」を 2006 年より開始した [1–4]。PodCastle は, 有用な Web サービスを運用して音声認識技術の実用化と普及をはかりつつ, Web 上の膨大なデータ, 更には Web サービスを通じて形成される知識 (集合知) を活用することで, 音声認識の高度化を行うことを目的としている。また, NICT では, Web 上の膨大なデータや情報インフラを利活用する音声言語研究プロジェクト「MASTAR」が 2008 年より立ち上げられ, 産学官で連携して進められている [5]。更に最近では, スマートフォン等の携帯デバイスの普及により, 音声を利用したアプリケーションが浸透し, 音声認識技術が一般のエンドユーザにとってより身近な存在になるなど, Web を通じて音声認識技術が盛り上がりを見せている。

* Automatic speech recognition vitalized by the Web.

** Jun Ogata and Masataka Goto (National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, 305–8568)

2. Web で生きる音声認識

最近, Web 上で様々な音声アプリケーションやサービスが展開され, 音声認識技術を多くのエンドユーザが利用できるようになってきている。ここでは, そうした「Web で**生きる**音声認識」に関する研究事例, 応用事例について述べる。

2.1 音声情報検索

音声情報検索 (音声ドキュメント検索) は, 音声認識技術の有力な応用の一つとして活発に研究がなされてきたが, Web 上で展開された例はそれほど多くはない。しかし, 最近では動画共有サイトやポッドキャストが急速に普及したことで, 動画等の「音声」を含むコンテンツ (便宜上, 以下では単に**音声コンテンツ**と呼ぶ) が大量に集積されるようになった。そのため, Web 上の音声コンテンツに対する情報検索や索引付け, 書き起こしへのニーズが高まっている。

音声情報検索システムの先駆的な事例[1]としては, 音声認識技術による索引付けにより Web 上の特定のラジオ番組のコンテンツを検索可能とした「SpeechBot」[6] がある。その後, Web 上の音声コンテンツとして普及したポッドキャストの検索を行う Web サービスとして, TVEyes 社の「Podscope」, BBN 社の音声認識・検索技術をベースとした「EveryZing」が公開された[2]。

我々は, Web 上の膨大な音声コンテンツを全文検索可能な Web サービス「PodCastle」を 2006 年 12 月に公開, 運用をしている (http://podcastle.jp)。PodCastle では大語彙連続音声認識技術により音声コンテンツをテキスト化 [7], それを索引情報とすることで, Web ブラウザ上で全文検索ができる。公開当初は, 日本語ポッドキャストを対象

[1] ここでは主として Web 上の音声コンテンツを対象としたシステム・サービスを取り上げる。

[2] ただし, 現在はいずれも無料での公開は終了し, 商用のシステムとなっている。

としていたが，その後，YouTube，Ustream，ニコニコ動画といった動画共有サイトのコンテンツにも対応した [4]。更に Web サービスの多言語対応も進めており，2011 年 10 月にはその第一弾として，エジンバラ大学音声技術研究所（CSTR）から音声認識に関する協力を得て英語版（英語の音声コンテンツの音声認識・全文検索）を公開した（http://en.podcastle.jp）。日本語版は産総研で研究開発した音声認識システムを用いているが，英語版は CSTR が中核研究機関として実施した欧州連合（EU）の研究枠組み計画（FP6 AMI 及び FP6 AMIDA）で開発され，同研究所が PodCastle 用に運用している音声認識システムを用いている。PodCastle では，ユーザが検索クエリとしてキーワードをタイプ入力すると，音声コンテンツ中で該当する箇所をリストアップし，再生できるだけでなく，音声コンテンツに対する音声認識結果もテキストとして閲覧可能となっている。そして最も特徴的な点として，Web ブラウザ上でユーザが容易に音声認識誤りを訂正できる機能を備えており，音声コンテンツに対する**書き起こし**を作成することが可能となっている（訂正機能の詳細は 2.3 節で述べる）。

産総研の李らは，音声認識による動画検索 Web サービス「VOISER」を公開している。これは，音声コンテンツに対してサブワード単位（音素片）の音声認識を行って索引付けし，サブワード単位でのマッチング手法に基づき検索を行う [8]。そのため言語に依存しない検索システムとなっているのが特長である。また，ユーザが Web ブラウザ上で検索を行う際は，テキストだけでなく，音声によるクエリ入力が可能である。

Google では，YouTube 上の米大統領選に関連した動画を対象に，音声認識による索引付け，検索可能なサービス「Google Audio Indexing」が試験的に公開された。現在では，YouTube 全般の動画中に字幕を表示する「自動キャプション機能」として，音声認識技術が応用されている（検索については未対応）。

2.2 音声入力 Web アプリケーション

Web を介してユーザが音声認識技術を利用できる一つの手段として，西村らは Web アプリケーション「w3voice」を公開している [9]。w3voice は，個々の Web システムに対して，事前の特別なプログラムのインストールなしで音声による入力インタフェースを容易に追加するもので，音声入力機能つきの様々な Web サイトを構築・拡張することができる。

2.3 携帯デバイスアプリケーション

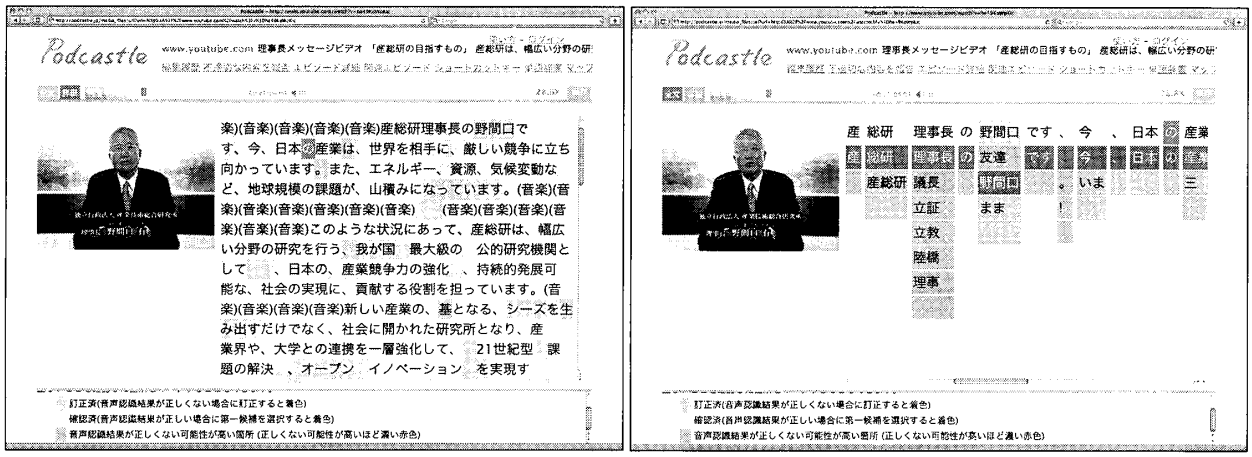
スマートフォン等の携帯デバイスの急速な普及は，一般のエンドユーザにとって音声認識技術を更に身近なものにしている。場所を選ばず Web に接続可能であり，キーボード等の入力デバイスの利用が難しい携帯デバイスにおいては，音声は非常に有用な入力手段になり得る。携帯デバイス上の音声入力を想定したタスクの一つとしてボイスサーチ [10] があり，音声認識分野におけるホットな研究課題となっている。ボイスサーチのアプリケーションとしては，Google による「Google Search by Voice（Google 音声検索）」や Yahoo!による「音声検索」がリリースされ話題となった。音声翻訳も携帯デバイスによって利用が広がったタスクであり，NICT による「VoiceTra」や Google による「Google 翻訳」など多くのアプリケーションがリリースされた。最近では，Apple の「Siri」や NTT ドコモの「しゃべってコンシェル」といった，検索だけでなく簡単なスマートフォン上の操作や軽い対話も行える秘書機能アプリケーションが登場し，人気を博している。

3. Web で活きる音声認識

近年，音声認識研究がより実世界・実環境の音声データを扱うようになった結果，Web という膨大かつ多様な知識を音声認識システムに活用する研究が活発に行われるようになった。特に，前述のような Web 上のアプリケーションやサービスは，Web という情報インフラを通じて，幅広い多くのユーザに利用されるというメリットがある一方で，バックグラウンドの音声認識技術には総じて高い性能，頑健性が要求される。そのため，既存のコーパスやデータベースだけではその要求に答えることは困難である。ここでは，Web 上のデータや知識を活用して性能向上をはかる「Web で**活きる**音声認識」に関する研究動向について述べる。

3.1 オフラインの音声コンテンツに対する音声認識

Web 上の様々な知識源の中でも最も手軽に利用でき，音声認識に有用と考えられるものとして，第



全文モード　　詳細モード

図–1 PodCastle の音声訂正インタフェース（全文モードと詳細モードはコンテンツ再生中でも切り替え可能）

一にはテキストデータが挙げられる。音声認識の言語モデリングにおける Web 上のテキストデータの活用は，2000 年以降多くの試みが行われている。当然ながら Web 上のテキストを何でも使えばいいわけでなく，多くの方法ではテキスト検索エンジンを利用することで，認識に有用なテキストのみを収集する。例えば，特定のドメイン（話題）に関連したテキストを収集したり [11]，会話特有のフレーズをクエリとして検索し，会話スタイルテキストを収集する [12] ことで，音声認識用の言語モデルを構築する。こうした静的な言語モデリングに対し，Web を利用することで認識対象の音声やドメインに動的に適応した言語モデルを構築する試みもなされている [13, 14]。

3.2 Web 上の音声コンテンツに対する音声認識

前節で紹介した研究事例では，主として Web 上にはないオフラインのコンテンツの音声認識を目的としたものであるが，それに対し，最近では Web 上の音声コンテンツの重要性も高まっている。

PodCastle においては，音声認識の対象が Web 上の動画，ポッドキャストといった話題が多種多様な音声データとなる。そのため，言語モデルとしてはできるだけ幅広い話題や語彙をカバーするだけでなく，日々移り変わる最新の話題に対応できなければならない。そこで PodCastle では，Yahoo! ニュース[3]に代表されるニュースアグリゲーション Web サイトにおける膨大なテキスト記事（Web ニューステキスト）を活用し，言語モデルを構築している [7]。重要な点として，Web ニューステキストを利用することで，最新の話題や語彙を日々学習し，言語モデルのアップデートを自動的に行っている。また，別の Web 上の知識源として Web キーワード辞書（はてなキーワード）を利用することで，形態素解析辞書の更新や，新出単語の読みの自動獲得も行っている。

[3] http://headlines.yahoo.co.jp/hl

4. Web で生きる中で Web を活かす音声認識

ここまで主として，Web を「テキストや音声といった多様で膨大なデータが蓄積する情報源」として音声認識研究に活用した研究事例に着目してきた。Web を活用することで音声認識の性能向上は得られるものの，2 章で述べたような Web サービス・アプリケーションを構築・運用したり，Web 上の多様なコンテンツを扱う上では課題は多く，さらなるブレークスルーが必要となる。そこで注目されるのが，Web 上の不特定多数のエンドユーザ（ここでは音声認識利用者）による協力や貢献により形成される知識源，すなわち**集合知（wisdom of crowds）**の活用である。ここでは，Web サービス・アプリケーションを通じて集合知を活用していく，いわば「Web で生きる中で Web を活かす音声認識」に関する研究動向について述べる。

4.1 携帯デバイスアプリケーションにおける事例

ボイスサーチは，音声入力により Web 検索ができる有用なアプリケーションであると同時に，

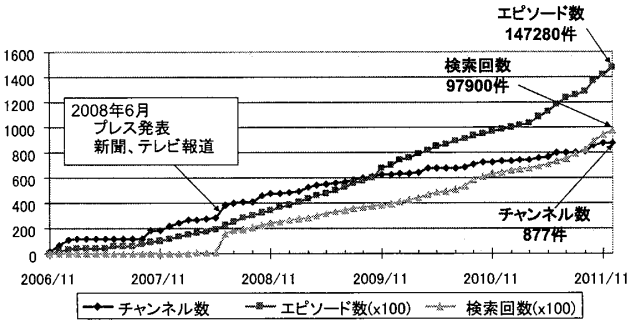
（システム側にとっては）音声データを収集するためのインタフェースとしてみることができる。スマートフォン等の携帯デバイスを通じて，不特定多数のエンドユーザからの音声データやその他情報を大規模に蓄積し[4]，それらを利用して音声認識システムを洗練させ，ボイスサーチの性能やユーザビリティを向上させることができる。

文献 [15] では，Google Search by Voice における音響モデルの性能改善について報告されており，ユーザからの入力音声 1,000 時間（＋人手による書き起こし）を学習に利用することで，ベースとなる音響モデル（前身サービスである GOOG-411 で構築されたモデル）に比べて格段に大きな向上が得られている。また，5,000 時間の音声データを教師なし学習することでさらなる性能改善が得られたことも報告されている。

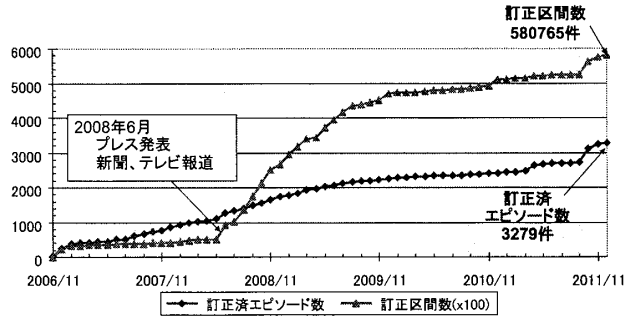
4.2 PodCastle における集合知活用

PodCastle では，Web 上の音声コンテンツの書き起こしを生成し，更には音声認識の性能向上をはかるために，ユーザからより積極的な協力（アノテーション，ここでは音声認識誤りの訂正）を得る仕組みを構築した。PodCastle では，図–1 に示すように競合候補のリストという形で訂正インタフェース [16, 17] を提供し，ユーザは Web ブラウザ上で構築された本インタフェースを通じて，認識誤りが見つかれば，「候補選択」，「タイプ入力」のいずれかの手段で訂正を行うことができる。PodCastle は，こうした不特定多数のユーザからの訂正情報を Web 上のコンテンツの実音声データとともに蓄積していくことで，音声認識システムの学習を行っていく。ここで，PodCastle におけるこれまでの利用状況を図–2, 3 に示す。2012 年 1 月 18 日時点で，登録済のエピソード（ファイル単位の動画あるいは音声データ）数は 150,003 件であり，そのうち 3,299 件に対して総計 583,799 箇所の区間（単語）がボランティアベースにより訂正されている。PodCastle では，こうした不特定多数のユーザからの協力を活用して，音響・言語モデルを日々学習し，音声認識・検索システムへの反映をすべて自動的に行う仕組みを構築し，運用している [7]。

[4]何をどこまで「集合知」と捉えるかは議論の余地があるが，ここでは Web を通じてユーザから入力された音声データや利用履歴なども含めて議論する。



図–2 PodCastle の利用状況：登録済のチャンネル数，エピソード数，検索回数の累積回数



図–3 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

話題や収録環境が多種多様なコンテンツに対して，いかに音響・言語モデルの学習を行うかは難しい問題であるが，PodCastle では「チャンネル」と呼ぶ単位ごとに学習を行い，チャンネル別音響・言語モデルを構築している。ここでチャンネルとは，対象としているサービス（YouTube，Ustream，ニコニコ動画，ポッドキャスト）それぞれで定義されている，個々のコンテンツがまとめられた単位[5]である。このようなチャンネルは，話者，話題，収録環境等，認識性能に影響を及ぼす条件が比較的類似する傾向があるため，モデル学習の際には有効に働くと考えられる。人気のあるチャンネルほどより多くの訂正が行われるため，我々のモデル学習の仕組みでそうしたチャンネルの学習効果が高まって，精度の高い音響・言語モデルが構築されることが期待できる。実際に，集積された訂正情報をもとに，音響・言語モデルを学習して認識性能を調査したところ性能が大きく改善し，ユーザからの訂正により入力されたチャンネル特有の語彙やフレーズ等も認識できることが確認された [18–21]。

[5]ニコニコ動画における「マイリスト」なども一つのチャンネルとして定義される。また，ポッドキャスト（RSS＋音声ファイル群）もここでは一つのチャンネルと呼ぶ。

4.3 音声研究におけるクラウドソーシング

近年，Webを通じて不特定多数の人々に業務委託（アウトソーシング）を行う，いわゆる「クラウドソーシング[6]」がWebにおけるトレンドの一つとなっている。音声研究においても，コーパス構築やシステム評価といった部分にクラウドソーシングを活用する試みが増えており，Interspeech 2011では音声研究におけるクラウドソーシングに関するスペシャルセッションも設けられた[22]。こうした試みの原動力は，Amazon Mechanical Turk（MTurk）に代表されるクラウドソーシングのためのプラットフォーム，Webサービスが登場したことであり，こういった有償ではあるが安価なサービスにより，研究開発におけるコスト削減，効率化が実現されている。これまでの報告によれば，クラウドソーシングの適用事例は，(1) 音声データ収集，(2) 音声データラベリング・書き起こし，(3) システム性能評価，の大きく3種類に分類でき，現状ではほとんどの研究事例でMTurkが利用されている[22]。PodCastleは，上記(2)の音声データ書き起こしをボランティアベースで行っており，クラウドソーシングに基づく音声研究の世界初の事例と位置づけることができる。MTurkのようなサービスにより低コストで効率的に集合知を活用し，様々な大規模データを形成できることは，「音声研究における集合知活用」の土壌を広げる意味でも意義が大きい。

5. おわりに

本解説では，現在Web上で発展している音声認識技術についての研究動向について述べた。音声研究でWebを利用するメリットとして，一つは集積する膨大なデータの利用が挙げられるが，今後は4章で述べたような「集合知」を形成し，研究開発に利用することがますます重要になると考えられる。

MTurkに代表される，金銭的報酬を与えて不特定多数に仕事を依頼するクラウドソーシング型Webサービスの登場は，低コストで効率的に集合知を形成し利用できるため，今後の研究開発の手段や体制そのものにも変革をもたらす可能性がある。しかし，こうしたアプローチでは十分に高い品質のアノテーションが得られたものの，いたずら（荒らし）対策が不可欠で，必要とされるアノテーションの量に比例して，人間の労力と対価が増えるといった問題も存在する。その問題を解決する一つのアプローチとして，ゲームとして楽しませながらアノテーションを付与させる方法が試みられている[23, 24]。

それに対し，我々が推進しているPodCastleでは，ユーザは貢献するとサービスが改善して自分を含む他のユーザの役に立てるということを明確に意識できるので，貢献しようというより強い動機に基づいてアノテーション（訂正）可能な点が特徴的である。こうしたアプローチの新たな事例として，我々は，不特定多数のユーザが協力しながら，協調的に音声コンテンツの書き起こしを行う「Yourscribe」を提案し，開発を進めている[25]。

アノテーション，そしてそれに基づく学習といったような，ユーザと音声認識技術との「インタラクション」をWeb上で実現・展開することは，音声認識技術の社会への浸透，普及にもつながっていく。今後そうした研究事例が増え，音声認識研究がますます盛り上がっていくことを期待したい。

[6] 日本語では同じ「クラウド」でも，冒頭でも述べたクラウドコンピューティング（**cloud** computing）はWebを通じて**計算リソース**を利用するものであり，クラウドソーシング（**crowd** sourcing）はWebを通じて**人的リソース**を活用するものである。

文　献

[1] 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).

[2] M. Goto, J. Ogata and K. Eto, "PodCastle: A Web 2.0 approach to speech recognition research," *Proc. Interspeech 2007*, pp. 2397–2400 (2007).

[3] 後藤真孝, 緒方　淳, 江渡浩一郎, "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," 人工知能学会誌, **25**, 104–113 (2010).

[4] M. Goto and J. Ogata, "PodCastle: Recent advances of a spoken document retrieval service improved by anonymous user contributions," *Proc. Interspeech 2011*, pp. 3073–3076 (2011).

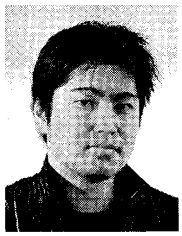
[5] 中村　哲, 清水　徹, 柏岡秀樹, 鳥澤健太郎, 隅田英一郎, "音声・言語研究拠点MASTARプロジェクトについて," 音講論集, pp. 63–64 (2009.3).

[6] J.-M.V. Thong, P.J. Moreno, B. Logan, B. Fidler, K. Maffey and M. Moores, "Speechbot: An experimental speech-based search engine for multimedia content on the web," *IEEE Trans. Multimedia*, **4**, 88–96 (2002).

[7] J. Ogata, M. Goto and K. Eto, "Automatic Transcription for a Web 2.0 Service to Search Podcasts," *Proc. Interspeech 2007*, pp. 2617–2620 (2007).

[8] 岩田耕平, 伊藤慶明, 小嶋和徳, 石亀昌明, 田中和世, 李　時旭, "語彙フリー音声文書検索手法における新しいサブワードモデルとサブワード音響距離の有効性の検証,"

情報処理学会論文誌, 48, 1990–2000 (2007).
[9] 西村竜一, 三宅純平, 河原英紀, 入野俊夫, “音声入力・認識機能を有する Web システム w3voice の開発と運用,” 情処研報音声言語情報処理, 2007-SLP-68-3, pp. 13–18 (2007).
[10] Y.-Y. Wang, D. Yu, Y.-C. Ju and A. Acero, “An introduction to voice search,” *IEEE Signal Process. Mag.*, 25, 28–38 (2008).
[11] R. Nishimura, K. Komatsu, Y. Kuroda, K. Nagatomo, A. Lee, H. Saruwatari and K. Shikano, “Automatic n-gram language model creation from web resources,” *Proc. Eurospeech 2001* (2001).
[12] I. Bulyko, M. Ostendorf and A. Stolcke, “Getting more mileage from Web text sources for conversational speech language modeling using class-dependent mixtures,” *Proc. HLT-NAACL 2003*, pp. 7–9 (2003).
[13] A. Berger and R. Miller, “Just-in-time language modeling,” *Proc. ICASSP '98*, pp. 677–680 (1998).
[14] S. Meng, K. Thambiratnam, Y. Lin, L. Wang, G. Li and F. Seide, “Vocabulary and language model adaptation using just one speech file,” *Proc. ICASSP 2010* (2010).
[15] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Garret and B. Strope, “Google search by voice: A case study,” in *Visions of Speech: Exploring New Voice Apps in Mobile Environments, Call Centers and Clinics*, A. Neustein, Ed. (Springer, Heidelberg, 2010).
[16] J. Ogata and M. Goto, “Speech repair: Quick error correction just by using selection operation for speech input interfaces,” *Proc. Interspeech 2005*, pp. 133–136 (2005).
[17] 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, 48, 375–385 (2007).
[18] 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論集, pp. 91–96 (2009).
[19] 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言語モデル学習,” 情処研報音声言語情報処理, 2010-SLP-80-10 (2010).
[20] J. Ogata and M. Goto, “PodCastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription,” *Proc. Interspeech 2009*, pp. 1491–1494 (2009).
[21] J. Ogata and M. Goto, “PodCastle: Collaborative training of language models on the basis of wisdom of crowds,” *Proc. Interspeech 2012* (2012).
[22] G. Parent and M. Eskenazi, “Speaking to the Crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges,” *Proc. Interspeech 2011*, pp. 3037–3040 (2011).
[23] L. von Ahn, “Games With A Purpose,” *IEEE Comput. Mag.*, 39, 92–94 (2006).
[24] S. Luz, M. Masoodian and B. Rogers, “Supporting collaborative transcription of recorded speech with a 3D game interface,” *Knowledge-based Intell. Inf. Eng. Syst.*, 6279, 394–401 (2010).
[25] 浮田俊輔, 緒方 淳, 後藤真孝, 小林哲則, “ライブストリーミングのための協調的音声書き起こしシステム,” 情処研報音声言語情報処理, 2011-SLP-85-8 (2011).



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，音声情報検索，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。

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）。現在，産業技術総合研究所 情報技術研究部門 上席研究員 兼メディアインタラクション研究グループ長。統計数理研究所 客員教授，筑波大学大学院 准教授（連携大学院），IPA 未踏 IT 人材発掘・育成事業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，29 件受賞。