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

2007－ＨＩ－122 (8)
2007－SLP－ 65 (8)
2007／2／9

PodCastleの実現: Web 2.0に基づく音声認識性能の向上について

緒方 淳　　後藤 真孝　　江渡 浩一郎

産業技術総合研究所

jun.ogata [at] aist.go.jp

あらまし　本稿では，ポッドキャストを検索できるWebサービス「PodCastle」を実現するための音声認識手法について述べる．ポッドキャストでは多様な内容が異なる環境で録音されており，多数の未知語を含む新たな話題も多いため，従来の音声認識システムで適切に認識するのは困難だった．この問題を解決するために，本研究では，Web 2.0によって得られる様々なデータを用いることによって，継続的に，音声認識システムを改善していく．具体的には，各ポッドキャストの内容に応じた言語モデルの話題適応，Web 2.0のサービスを通じた単語発音の自動獲得，PodCastle上でユーザが音声認識誤りを訂正した結果を用いた未知語の学習等を試みた．実際にポッドキャストを対象とした認識実験を行い，性能向上に有効であることを確認した．

PodCastle: Techniques for Improving Speech Recognition Performance on the Basis of Web 2.0

Jun Ogata　　Masataka Goto　　Kouichirou Eto

National Institute of Advanced Industrial Science and Technology (AIST)

Abstract　This paper describes speech recognition techniques that enable a web service *"PodCastle"* for searching podcasts. Most previous speech recognizers had difficulties dealing with podcasts because they include various contents recorded in different conditions and new topics with many out-of-vocabulary words. To overcome such difficulties, we continuously improve speech recognizers by using information aggregated on the basis of Web 2.0. For example, the language model is adapted to a topic of the target podcast on the fly, the pronounciation of unknown words is obtained from a Web 2.0 service, and out-of-vocabulary words are automatically acquired by analyzing user corrections of speech recognition errors on PodCastle. The experiments we report in this paper show that our techniques produce promising results for podcasts.

1　はじめに

音声情報検索は，音声認識技術のアプリケーションの1つとして重要視され，近年でも活発に研究が展開されている [1]．しかしながら，現状の音声認識技術では，あらゆる音声データから検索に必要な索引情報(テキストやキーワード等)を，精度よく抽出することが困難なこともあり，Google等の代表されるテキストの検索のように日常的に利用されるには至っていない．一方，最近では，音声版のブログ(Weblog)ともいえる「ポッドキャスト」が普及し，Web上の音声データとして多数公開されるようになったため，そうした音声データに対する検索の重要性がより一層増してきたといえる．

そこで我々は，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はなく，詳細な閲覧，編集も可能なソーシャルアノテーションシステム「PodCastle[2]」の開発を行っている．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，システムが持つすべての情報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(アノテーション)する協力をしてもらうことで，音声認識率をシステムの運用中に向上させる枠組みを採用している．こうすることで，検索サービスとしての質を向上させるだけでなく，音声認識技術の底上げをはかることも狙っている．

本稿では，以上述べたPodCastleを実現するための音声認識手法について検討する．ポッドキャストは，その発話内容や録音環境などが多種多様であり，従来のタスクを限定した場合の音声認識に比べて多くの問題を含んでいる．本研究では，このような問題に対し，従来のように研究技術の枠組みだけで解決するのではなく，Web 2.0を通じて得られる知識やデータを積極的に利用することで，音声認識性能を改善するアプローチを採用する．具体的には，本稿では，多岐にわたる話題を認識するための言語モデル適応，Webからの単語発音の自動獲得，ソーシャルアノテーションシステムを通じて集めた，ユーザの訂正結果を利用した音響モデル学習などを試みる．以下，2章にてPodCastle音声認識システムの基本構成について説明し，3章ではWeb 2.0に基づいて音声認識性能を改善する手法について議論する．最後に，4章で本稿のまとめを行うとともに，本研究の意義について述べる．

2 PodCastle音声認識システムの基本構成

PodCastleにおいて最も特徴的なのは，サーバにアクセスしているユーザが誰でも，編集機能により，認識誤りの訂正を行える点である．訂正の際には，我々が以前に提案した「音声訂正 [3]」インタフェースを用いている．本インタフェースでは，従来の音声認識のように1つの単語列だけではなく，図1に示すように，複数候補を各区間ごとにまとめた「競合候補」のリストを提示する．そして競合候補の中に本来の正解があれば，ユーザはその単語を「選択」することで，容易に認識誤りを訂正できる．したがって，本音声認識システムにおいては，最終的な認識結果として，このような複数候補を出力する．以下では，ベースラインの大語彙連続音声認識システムの各要素技術について述べる．

2.1 音響イベント検出，発話区間推定

高精度な音声認識を実現するためには，認識すべき発話区間を正しく推定することが重要である．一般

図 1: 音声訂正インタフェースを備えたPodCastleの画面表示例

的な音声認識ソフトウェアや，講演音声などのように比較的雑音が少ないタスクにおいては，単純に音響パワーやゼロクロスを閾値処理することで．特定の音声区間を求めることが可能であるが，ポッドキャストのように様々な音響イベントが混在するタスクにおいてはそのような単純な手法を適用することはできない．

本研究では音声認識の前処理として，GMMを用いた音響イベント検出を行い，その結果に基づき，認識すべき発話区間を推定する．純粋な音声以外に，発生する様々な雑音に対してもモデル化を行うことが理想的であるが，ここでは第一段階として，音声，音楽，無音の3種類のみを音響イベントとして定義した．各モデルの学習データとしては，音声，無音モデルには新聞記事読み上げコーパス (JNAS) を，音楽モデルにはRWC研究用音楽データベース (RWC-MDB-P-2001，RWC-MDB-R-2001，RWC-MDB-J-2001) [4] (ただし，通常とは異なる歌声を含まないカラオケ版) を用いた．各モデルの混合数は64とした．

2.2 音響モデル

音響モデルには，単語間の調音結合も考慮した状態共有型triphoneモデルを用いた．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．本研究では，雑音対処の1つとして，雑音環境下音声認識において幅広く利用されているETSI Advanced Front-End[5]を音響分析に適用した．そして，最終的に，12次元のMFCCとパワー，およびそれぞれの Δ，$\Delta\Delta$ を算出し，計39次元の音響特徴量を求めた．学習データには，日本語話し言葉コーパス (CSJ) を用いた．triphoneモデルの状態数は，MDL基準を用いた状態クラスタリングにより自動決定した4513であり，1状態あたりの混合数は16とし

た．以下ではこれをベースライン音響モデルとする．

ポッドキャストにおいては，音楽などの非定常雑音も多く含まれるため，上記のような雑音抑圧処理を行ったとしても，ベースライン音響モデルとの間にミスマッチが生じ，十分な性能が得られない可能性がある．そこで，ポッドキャストの各エピソードごとにバッチ型の教師なし適応を行うことで，環境，話者等に対するミスマッチを軽減させる．まず，ベースライン音響モデルを用いて，語彙に依存しない連続音節認識を行い，適応処理に用いる教師信号 (音素ラベル) を得る．そして，MLLR[6] により，音響モデルのパラメータ (平均，分散) を更新する．

2.3 言語モデル

ポッドキャストのように，幅広いタスクやドメインの音声を扱う場合には，言語モデルでカバーすべき話題や語彙を事前に絞り込むことはできない．したがって，ベースライン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，比較的大規模なモデルを構築する必要がある．本研究では，基本的な学習用コーパスとして，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利用した．後者のコーパスは，ポッドキャスト音声データにおいて頻出する自然発話に対応するために学習に含めている．

しかしポッドキャストの場合，最近の話題や語彙を含むものが多く，事前にいかに大規模な言語モデルを用意したとしても，未知語の問題が劇的に解決されることはない．そこで，本研究では，日々更新されている Web 上のニュースサイトのテキストを，言語モデルの学習に利用して，言語モデルの性能を改善する．具体的には，総合的な日本語ニュースサイトである Google ニュースと Yahoo!ニュースに掲載された記事のテキストを，それぞれのサイトのカテゴリごとに分類して毎日収集し，それらを言語モデルの学習に利用する．以下の実験においては，2006 年 8 月 1 日から 2006 年 12 月 31 日までの間に収集したテキストデータを用いた．以上の 3 種類のテキストコーパス (新聞記事，日本語話し言葉コーパス，Web ニュース) を用いて，最終的に語彙数 152163 の trigram モデルを構築した．

2.4 デコーディング

本音声認識システムのデコーディングは段階的探索に基づいている．まず，bigram を用いた N-best デコーディングにより単語グラフを生成する．このときの探索アルゴリズムとしては，back-off 制約 N-best 探索法 [7] を用いている．次に，trigram を用いて，生成された単語グラフを，trigram 制約の単語グラフに拡張する．最後に，trigram 制約の単語グラフに対して，consensus デコーディング [8] を行い，confusion network を生成する．なお，最終的に出力された confusion network は，図 1 に示す訂正インタフェースに利用される．

表 1: 評価用音声データ

ID	title	#episode
A	森永卓郎 経済コラム	3
B	森本毅朗・スタンバイ！	2
C	読売ニュース ポッド …	3
D	報知芸能ポッドキャスト	2

表 2: テストセットパープレキシティ(括弧内の数値は未知語数)

ID	LM	LM+web
A	98.5 (9)	89.9 (4)
B	84.0 (0)	76.2 (0)
C	81.5 (69)	57.0 (39)
D	146.6 (8)	99.6 (5)

2.5 ベースラインシステムの性能評価

実際にポッドキャストを対象としてベースラインシステムの性能評価を行った．表 1 に実験で用いた音声データを示す．ここで，“ID” はポッドキャストの番号，“title” はポッドキャストのタイトル，“episode” は本データベースに登録されているエピソードの数である．発話スタイルとしては，A，B は講演音声に近く，自発的な音声といえる．一方 C，D は読み上げニュース音声である．また，D に限っては常に背景に音楽が流れている．

まず言語モデルについての評価を行う．各言語モデルのテストセットパープレキシティと未知語数を表 2 に示す．表中 “LM” は新聞記事と CSJ から学習したモデルであり，“LM+web” は Web ニュースのテキストも学習に含めたモデルである．全体的に Web ニュースを利用することで，パープレキシティ，未知語数ともに削減できていた．とくにニュース音声 (C, D) に対する改善が大きいことがわかる．これは Web ニュースによって最新の話題，語彙にうまく対応できたことを表している．ポッドキャストごとの認識率を表 3 に示す．ここで表中の括弧内の数値は，confusion

表 3: 各言語モデルにおける認識率 (括弧内は NAC)

ID	LM	LM+web
A	73.3% (90.3%)	74.0% (91.0%)
B	69.3% (85.3%)	70.2% (88.1%)
C	75.1% (88.2%)	81.9% (95.5%)
D	49.1% (73.6%)	60.7% (86.3%)

network の精度 (NAC: network accuracy) であり，これは，confusion network 中の，正解に最も適合する単語列に対する認識率を表している．パープレキシティと同様，認識率においても Web ニュースを利用することの効果が確認された．

以上の結果からも，ポッドキャストのように，日々発信され，最新の話題や語彙が頻出する音声を扱うには，言語モデルがそれらに対応できるように，日々アップデートしていくことが重要であると考えられる．

3 Web 2.0 に基づく音声認識性能の改善

Web 2.0 を利用して得られる知識やデータを利用した，音声認識性能の改善手法について述べる．ここで「Web 2.0 を利用して得られる知識やデータ」とは，PodCastle 自身が持つ Web 2.0 的な機能やアーキテクチャを通して得られるものだけに留まらず，PodCastle 以外の，世の中にある様々な Web 2.0 的サービスにより得られるものも包含している．Web 2.0 的サービス，サイトにおいては，「集合知」あるいは「参加型アーキテクチャ」という考え方に基づき，不特定多数のユーザからの貢献により，様々な知識やデータが集積されている．それらは日々更新され続け，結果として膨大な知識が形成されるものもある (例えば，Wikipedia[9] など)．

このように Web 2.0 を通じて生成される知識やデータの中には，音声認識システムを学習するために有用なものも存在すると考えられる．ここで重要なのは，学習に利用する知識やデータは日々更新され続けるため，それに合わせて音声認識システム側をアップデートすることで，日々音声認識性能を改善させることが可能になる点である．

以下では，本稿にて検討した，Web 2.0 に基づく音声認識性能の改善手法を順に説明する．

3.1 RSS メタ情報を用いた言語モデルの話題適応

ポッドキャストには，音声データとともに，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フォーマット RSS (Really Simple Syndication) が必ず付与されている．RSS とは，Web 2.0 の構成要素の 1 つであり，その中にはコンテンツに対する様々なメタ情報が記述されている．

ここでは，RSS から取得できるメタ情報を利用した，音声認識システムの学習方法について検討する．メタ情報中には，音声データに対するタイトルや要約も含まれており，本研究ではこれらのデータを利用した言語モデルの話題適応を行う．

話題適応はポッドキャスト中の各エピソードごとに行われる．まず，タイトル，要約のテキストデータから，キーワードを抽出する．次に，抽出したキーワードをクエリとして，テキスト検索エンジンにより Web ページを収集することで，音声データの話題に特化したテキストを取得する．取得したテキストで言語モデルを作成し，ベースとなる大規模な言語モデルとの間で線形補間を行うことで，最終的な言語モデルを生成する．

本手法の効果を確かめるべく，2.5 節の評価用データのうち，A，B，D の 3 つのポッドキャストの各エピソード (合計 8) を用いて，パープレキシティならびに未知語数に関する評価を行った．キーワード抽出は，要約のテキストから固有名詞のみを自動抽出することで行った．また，タイトルに関しては，全体を 1 つのキーワードとした．検索エンジンとしては Yahoo!API を使用し，取得するページ数は，1 回の検索において最大 200 とした．表 4 より，パープレキシティ，未知語ともに削減されたが，今回のデータではそれほどの効果は見られなかった．この理由としては，今回のデータに関しては，未知語は少なく，話題自体もベースとなる言語モデルにおいてある程度カバーされているからである．ただし，ポッドキャストにおいては，偏った話題のものや専門的なものも多く，そのような音声データに関してはさらに有効に働くと考えられる．

3.2 Web からの単語発音の自動獲得

音声認識システムにおいて，登録されている個々の単語の発音 (読み) をどのように設定するか (発音モデリング) は認識性能を左右する重要なポイントである．従来のように限定されたタスクにおける音声認識においては，語彙も限定されていることもあり，事前に手

表 4: テストセットパープレキシティ(括弧内の数値は未知語数)

ID	ベース	適応
A	89.9 (4)	83.5 (1)
B	76.2 (0)	73.1 (0)
D	99.6 (8)	96.4 (6)

動で発音を付与することが可能であった．あるいは，個々の読みも整備された汎用的な形態素辞書 (例えば [10]) によって，登録されている全単語の発音をカバーすることができた．

それに対し，ポッドキャストにおいては，日々新たな音声データが発信され，内容は多岐にわたり，しかも世の中の最新の動向や話題について話されていることも多く，汎用的な形態素辞書で全てをカバーすることは不可能である．特に，ローマ字表記の専門用語や造語など (例えば,「PodCastle」など) に関しては，正しい発音を自動的に付与することは困難である．

そこで，Web から単語の発音を自動的に取得することを考える．Web 2.0 的サービスの 1 つである「はてなダイアリーキーワード [11]」では，集合知によって，様々なジャンルのキーワードとそれらに対する説明文が整備されている．さらに，キーワードとともにそれに対する読み (ふりがな) までも定型のフォーマットにて記述されており，必要な単語に対する読みを容易に取得できる．これらは日々集積，更新されており，原稿執筆時点において約 18 万のキーワードが登録されている．このキーワード群には，新語，造語なども含めて，世間で話題となったキーワードは多くカバーされており，それらがポッドキャストにおいて話題にされることも頻繁にある．特に前節で述べた言語モデルの適応化の際には，それぞれの話題ごとに特化したキーワードが多く登録される可能性が高いため，有効に働くと考えられる．

3.3 ユーザの訂正結果に基づく音声認識システムの学習

ユーザの訂正結果は，音声認識システムを改善する上で重要である．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を提供しているため，様々な形式の訂正結果が得られると考えられる．例えば，全音声区間の発話内容を正確に再現したものもあれば，聞き取れた箇所だけ訂正したもの，あるいはその音声においてキーワードとなる箇所のみ訂正したものなどが挙げられる．以上のいずれの形式であれ，なんらかの形で音声認識システムの学習に生かし，ユーザの協力を反映していくことが重要であると考える．

以下では，本システムで検討している主な学習手法について順に述べる．

3.3.1 音響モデル

ユーザの訂正結果から，音声データに対する忠実な書き起こしが得られると，それらを用いることで，従来音声コーパスを用いて行っている場合と同様に，音響モデルの学習が可能である．このような書き起こしが大量に集まると，多種多様な音響特性を含む音声データベースが構築でき，それを学習に利用することでよりロバストかつ高精度な音響モデルを実現できる可能性がある．

一方，ユーザが聞き取れた箇所だけ訂正したときのように，音声データに忠実な書き起こしではない場合においても，近年検討されている「少量教師あり学習 (lightly supervised training)[12]」などを適用すればある程度の性能改善は期待できる．

3.3.2 言語モデル

言語モデルを学習するためには，音声に対して忠実な書き起こしが必要であると考えられる．現状で，認識が困難な複数人での会話音声などの書き起こしがある程度得られると，言語モデルの学習・適応が可能となり，認識性能の改善が期待できる．

3.3.3 未知語

ユーザの訂正結果から，未知語を認識システムの辞書に登録する方法である．本来の正解単語が，訂正インタフェースにおける競合候補中になく，ユーザがその区間にタイプ入力したとき，その単語が未知語であった場合に認識システム側の辞書に新たに登録する．登録した単語の発音は，誤認識した単語の区間に対する発音系列 (音素列) を，連続音素認識により自動的に求める．そして，得られた音素列を，新たに登録した単語の発音系列として登録する．

4 おわりに

本稿では，集合知を活用した音声情報検索用 Web サービス「PodCastle」を実現するための音声認識手法について検討した．PodCastle は,「音声認識研究 2.0[13]」という新たな研究アプローチを具現化したものであり，そこでは，Web サービスを中心として，日々

増え続ける多種多様な音声を認識していく必要がある．そのためには，本研究で提案したような，日々更新されるWebサイトを利用して最新の話題を常に認識可能にする手法や，Web 2.0に基づく話題適応，単語発音の自動獲得の手法等が重要となる．これらの一連の手法は，「音声認識システム自体が日々成長する」という新たなアプローチと言える．音声認識手法としての本研究の第一の意義は，音声認識研究1.0によって高性能化してきた現在の音声認識システムに加え，この「音声認識システム自体が日々成長する」アプローチがどこまで性能を向上できるかを探求することにある．

さらに，文献[13]でも述べたように，不特定多数のエンドユーザにアノテーション(認識誤り訂正)をしてもらうことで，ユーザに「音声認識を育ててもらう」というアプローチもPodCastleの重要な特長である．音声認識手法としての本研究の第二の意義は，この「音声認識を育ててもらう」ことを実現するために，認識誤りの訂正結果から音声認識性能の向上をどこまで引き出せるかを探求することにある．

このように音声認識が成長し，かつ，育ててもらうという2つの重要なアプローチは，音声認識研究2.0における音声認識システムを実現する上で，車の両輪のようにどちらも欠かせないものであり，今後もさらなる性能向上へ向けて研究を進めていく必要がある．今回の評価実験は小規模であったが，今後は，より大規模なデータセットを用いて評価実験を実施する予定である．また，現状では複数話者が会話をするフリートークのような音声データや背景音楽の含まれる音声データに対しては，著しく性能が低下する場合があることがわかっている．これらは従来の音声認識研究1.0でも認識されていた困難な課題ではあるが，Web 2.0の考え方を生かしながら，今後も様々なアプローチから改善へ向けて取り組んでいきたい．

謝辞

PodCastleの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（代表取締役 上津竜太郎 氏）と有限会社メロートーン（代表取締役 新井 俊一 氏）に感謝する．

参考文献

[1] 伊藤 克亘, 相川 清明, 秋葉 友良, 伊藤 慶明, 河原 達也, 南條浩輝, 西崎 博光, 安田 宜仁, 山下 洋一: “音声ドキュメント検索評価のためのテストコレクションの試作” 情処研報, 2006-SLP-64-24, pp.137-142, 2007.

[2] 緒方 淳, 後藤 真孝: “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.

[3] 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.

[4] 後藤 真孝, 橋口 博樹, 西村 拓一, 岡 隆一: “RWC研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース”, 情処学論, Vol.45, No.3, pp.728-738, 2004.

[5] ETSI ES 202 050 v1.1.1 STQ; ”Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms”. 2002.

[6] C.L.Leggetter and P.C.Woodland：“Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models”, Computer Speech and Language,Vol.9, pp.171-185, 1995.

[7] 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.84-D-II, No.12, pp.2489-2500, 2001.

[8] L.Mangu, E.Brill and A.Stolcke: “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Network” Computer Speech and Language,Vol.14, No.4, pp.373-400, 2000.

[9] Wikipedia: http://wikipedia.org

[10] 形態素解析システム茶筌: http://chasen.naist.jp/hiki/ChaSen/

[11] はてなダイアリーキーワード: http://d.hatena.ne.jp/keyword/

[12] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda: “Lightly Supervised and Unsupervised acoustic model training” Computer Speech and Language,Vol.16, pp.115-129, 2002.

[13] 後藤 真孝, 緒方 淳, 江渡 浩一郎: “PodCastleの提案: 音声認識研究2.0を目指して” 情処研報, 2007-SLP-65-7, 2007.