PodCastle: Web 2.0 に基づくポッドキャスト音声認識手法*

○緒方 淳，後藤 真孝，江渡 浩一郎 (産総研)

1　はじめに

我々は，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[1)2)3)]」の開発を行っている．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，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検討する．

2　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．PodCastle における訂正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 [4)]．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ことを目的とする．以下では，ベースラインの音声認識システムの各要素技術について述べる．

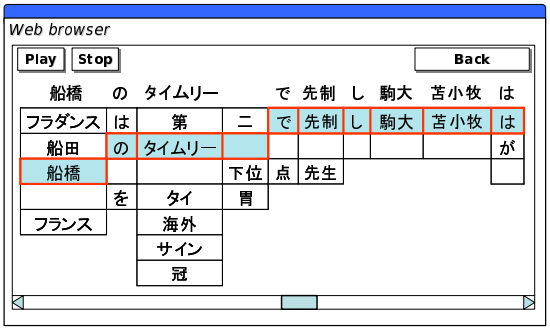

Fig. 1　PodCastle 訂正インタフェース

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．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．

2.2　音響モデル

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，Δ，ΔΔ 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．

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，静的に用意されたコーパスを利用して，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，Web 上のニュースサイトのテキストを学習に利用して言語モデルを日々アップデートすることで，最新の語彙や話題に対応できるようにする．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．

---

*PodCastle: Automatic Transcription of Podcasts on the Basis of Web 2.0.
by Jun Ogata, Masataka Goto, Kouichirou Eto (AIST)

Table 1　評価用音声データ

ID	内容	総時間 (sec.)	発話スタイル
A	経済コラム	192.1	自発
B	経済コラム	128.2	自発
C	ニュース	1159.6	読み上げ
D	ニュース	247.3	読み上げ
E	トーク (雑談)	486.6	自発

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以下に示すように段階的探索 (3-pass) に基づいている．

1. 高速な音素デコーディングを行い，得られた音素列を用いて音響モデルの教師なし適応を行う．適応手法としては MLLR を用いる．
2. 適応された音響モデルを用いて単語デコーディングを実行する．まず，bigram を用いた N-best 探索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 [6] を用いる．次に，trigram を用いて単語グラフをリスコアする．そしてその結果をもとに再度 MLLR 教師なし適応を実行する．
3. 適応された音響モデルを用いて，2. の単語デコーディングを実行し，単語グラフを再生成する．最後に，単語グラフに対して，consensus デコーディング [9] を行い，confusion network を生成する．

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．

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認識システムの性能評価を行った．なお，以降の実験では，我々の以前の報告 [7] とは異なる評価データを用いている (ただし文献 [8] とは同一)．

表 1 に実験で用いた音声データの特徴 (内容，総時間，発話スタイル) を，各ポッドキャスト (A,B,C,D,E) ごとに示す．本実験で用いたポッドキャストのエピソード☆1数はそれぞれ 1 である．これら 5 つのポッドキャストのうち，ニュース番組である C,D は，発話スタイルとしては読み上げ音声であるが，1 つのエピソード中に複数の話題が存在するという特徴がある．残りの A,B,E はいずれも自発的な発話スタイルであるが，経済コラムの A,B は講演音声に近く，事前にある程度のストーリーが頭に入った上での独話になっている．一方，E は 1 話者 (芸能人) によるトーク (雑談) 番組であり，より日常会話に近いくだけた発話スタイルとなっている．また，音響条件に関しては，A,B,C,E は雑音がほとんどないクリーンなデータであるが，D に限っては常に背景に音楽が流れている．

☆1 1 つのポッドキャストに含まれる一連の音声データ (MP3 ファイル) をエピソードと呼ぶ．

Table 2　パープレキシティ (ppl.) と未知語数 (#OOV)

ID	LM		LM+Web	
	ppl.	#OOV	ppl.	#OOV
A	112.2	4	104.1	2
B	86.8	0	72.6	0
C	86.5	10	62.1	10
D	146.6	8	99.6	5
E	281.0	18	266.2	18

Table 3　各言語モデルにおける認識性能

ID	LM		LM+Web	
	WER(%)	NER(%)	WER(%)	NER(%)
A	28.9	12.5	27.5	11.1
B	27.5	13.4	25.3	9.8
C	24.1	9.1	18.3	8.2
D	45.6	24.8	35.6	13.1
E	48.4	32.3	45.1	29.1

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されていた．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単語誤り率 (WER) を表 3 に示す．ここで表中の NER(Network Error Rate) は，confusion network の誤り率であり，confusion network 中の，正解に最も適合する単語列の WER を表している．パープレキシティと同様，認識性能においても 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手法

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[10] など)．

Table 4　パープレキシティ(ppl.) と未知語数 (#OOV)

	LM+Web		LM+Web+adapt	
ID	ppl.	#OOV	ppl.	#OOV
A	104.1	2	103.1	2
B	72.6	0	46.9	0
C	62.1	10	61.5	10
D	99.6	5	88.9	5
E	266.2	18	193.6	18

Table 5　各言語モデルにおける認識性能

	LM+Web		LM+Web+adapt	
ID	WER(%)	NER(%)	WER(%)	NER(%)
A	27.5	11.1	27.2	11.2
B	25.3	9.8	22.9	9.0
C	18.3	8.2	18.0	8.2
D	35.6	13.1	34.5	12.9
E	45.1	29.1	42.4	27.4

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，ブログ等で更新情報を通知するために用いられるフォーマット 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．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，ベースライン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．

実際に，2.5 節の評価用データを用いて本手法の評価を行った．キーワード抽出は，タイトル，要約のテキストを形態素解析し，その結果の中から固有名詞のみを抽出することで行った．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．結果を表 4，5 に示す．本手法の効果は，評価用データ中の各エピソードごとに様々であった．B,D,E においては，パープレキシティ，認識性能ともに向上がみられたことで，各エピソードごとの話題に対して関連のあるテキストが Web から収集できたといえる．一方，A,C においては，さほどの効果は得られなかった．これは，RSS 中のメタ情報として，エピソードの配信日時が記載されているだけで，話題に関する有益な情報が得られなかったためである．

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．従来のように語彙が限定されたタスクにおける音声認識においては，事前に手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 [11]) 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．実際に，前節までの実験で用いた言語モデル (LM+Web) 中の全単語のうち，少なくとも 11.5%(17438/152163) の単語に関しては本来の正しい発音を得ることはできていなかった．

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 [12]」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 20 万のキーワードが登録されている．実際に，はてなダイアリーキーワードを利用することで，上記の，正しい発音が得られなかった 17438 単語のうち，22.9%(3997/17438) の単語に対して正しい発音を与えることができた．残念ながら，今回の評価用データにはこれらの単語が出現しなかったため，認識率での効果を確認することはできなかったが，本手法で推定された単語は比較的ポピュラーなものが多く，様々なポッドキャストを認識する際に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認識システムの学習

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，本研究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，lightly supervised training[13] などを適用することで性能改善を期待できる．

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[2)3)]」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，本研究で検討した「音声認識システム自体が日々成長する」というアプローチが必要不可欠になると考えられる．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 [2)] 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ると考えられる．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，複数話者による会話音声や背景音楽の含まれる音声など，従来の音声認識研究 1.0 で判明していた困難な課題に対しても，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．

参 考 文 献

[1] 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.

[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, 2007-SLP-65-7, 2007.

[3] Goto, Ogata, Eto: PodCastle: A Web 2.0 Approach to Speech Recognition Research, *Proc. of Interspeech2007*, 2007.

[4] 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, Vol.48, No.1, pp.375-385, 2007.

[5] ETSI ES 202 050 v1.1.1 STQ

[6] 緒方, 有木: 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.

[7] 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, 2007-SLP-65-8, pp.41-46, 2007.

[8] Ogata, Goto, Eto: Automatic Transcription for a Web 2.0 Service to Search Podcasts, *Proc. of Interspeech2007*, 2007.

[9] Mangu, Brill and Stolcke: Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Network *Computer Speech and Language*,Vol.14, pp.373-400, 2000.

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

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

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

[13] Lamel, Gauvain and Adda: Lightly Supervised and Unsupervised acoustic model training, *Computer Speech and Language*,Vol.16, pp.115-129, 2002.