PodCastle: 集合知を活用した音響モデル学習による音声認識の性能向上 *

○緒方 淳，後藤 真孝 (産総研)

1 はじめに

我々は，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，運営を行っている．ポッドキャストは実環境の多様な音声データであり，従来の音声認識技術では高い認識率を達成することは難しい．そこで PodCastle では，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．

本研究では，上記の枠組みの一環として，PodCastle を通じて得られる集合知，すなわちユーザによる音声認識誤りの訂正結果を活用した音響モデル学習について検討し，ポッドキャスト音声認識において評価を行う．

2 ポッドキャスト音声認識

ポッドキャストは音声版のブログ (Weblog) ともいえる Web 上のコンテンツであり，個人の日記から大学の授業，ニュースまで多岐に渡る内容の音声データが日々配信されている．ポッドキャストは，一連の**エピソード**と呼ばれる音声データ (MP3 ファイルなど) 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 が付与されている (図 1)．

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．例えば，言語モデルの観点では，日々音声データが更新されることで発生する未知語や前述したような幅広いトピックを含む問題が挙げられる．一方，本稿で着目する音響モデルの観点では，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．また，ニュース，講演，雑談等，発話スタイルも多様である．PodCastle

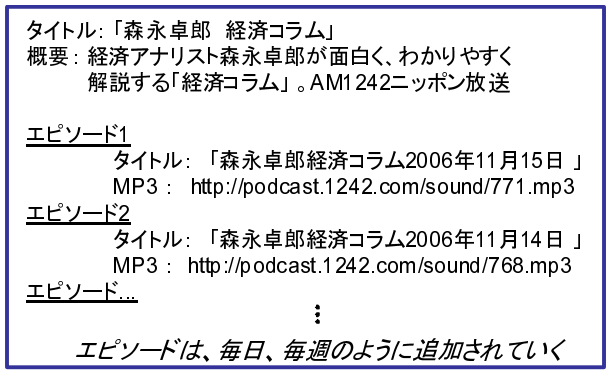

Fig. 1 ポッドキャストの構成 (RSS の例)

のこれまでの音声認識システム [3)] では，特徴抽出段階での雑音除去 [4)]，さらに個々のエピソードを認識する際にバッチ型の繰り返し教師なし適応 [3)] などによって対処を行っているが，音楽などの非定常雑音も多く存在することや，ベースラインが低いことで教師なし適応が十分に働かない等の影響で十分な性能とはいえなかった．

3 集合知を活用した音響モデル学習

ここでは，PodCastle を通じて実際に得られる不特定多数のユーザによる訂正結果，ポッドキャストの構成を利用した音響モデル学習法について述べる．

3.1 不特定多数のユーザによる音声認識誤りの訂正

PodCastle では，図 2 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 [5)] を提供している．ユーザは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．現在の状況として，サイトに登録されているいずれかの音声に対して，ほぼ毎日のペースで訂正がなされており，ポッドキャストによっては全エピソードのほぼ全ての認識誤りが訂正されているものもある．PodCastle の音声データに関する原稿執筆時点 (2009 年 1 月 20 日) での統計情報を表 1 に示す．ここで「訂正されたエピソード数」とは，１つのエピソード中で訂正が 1 カ所でも行われているものをカウントしている．

---

*PodCastle: Improvements of Speech Recognition by Using Acoustic Modeling Based on Wisdom of Crowds
by Jun Ogata, Masataka Goto (AIST)

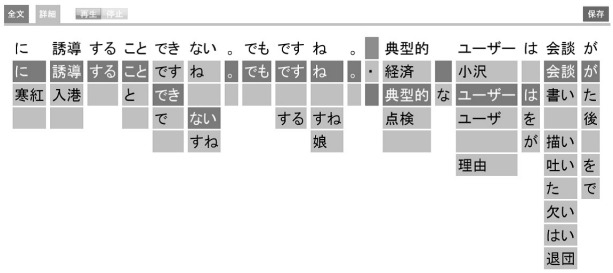

Fig. 2　訂正インタフェース

Table 1　PodCastle 統計情報

登録済みポッドキャスト数	482
エピソード (MP3 音声ファイル) 数	37825
訂正されたエピソード数	1489

3.2　ポッドキャスト依存音響モデル学習

従来のようにタスクが限定された音声認識の場合，そのタスクの特性に合った音声データを利用して単一の音響モデルが学習 (適応) されるが，ポッドキャストにおいては前述したように音声データ間の音響的な特性の差異が大きいために，精度の高い単一音響モデルを学習することは難しいと考えられる．

そこで，多様な音声データを音響的特性ごとにクラスタリングすることで，各特性ごとに音響モデルを構築する必要があるが，本研究では，第 1 ステップとして，PodCastle に登録されているポッドキャストごとに音響モデルの学習を行う．この理由として，同一ポッドキャスト中の各エピソードは，同じ音響的特性 (収録環境，話者，雑音，背景音楽，発話スタイル等) を持っている可能性が高いことが挙げられる．また，ポッドキャスト (RSS) の仕組みにより，認識対象となる各音声ファイルがどのポッドキャストのエピソードなのか，すなわち各音声ごとにどの音響モデルを認識時に適用すべきか自明であるという利点もある．以下，ポッドキャスト依存音響モデル学習の手順を説明する．

3.2.1　発話分割

各エピソードの音響ストリームを，音声の発話単位に自動分割する．本研究では，GMM を用いた音響イベントセグメンテーションを行い，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．ここで得られた各発話区間に対し，認識誤りの訂正により得られたテキストを割り当てる．PodCastle における訂正処理は，図 2 に示すように，音声認識結果 (confusion network[6]) 上の該当する区間で直接行われるため，このようなテキストの各発話区間への割り当ては自動的に正しく行うことができる．

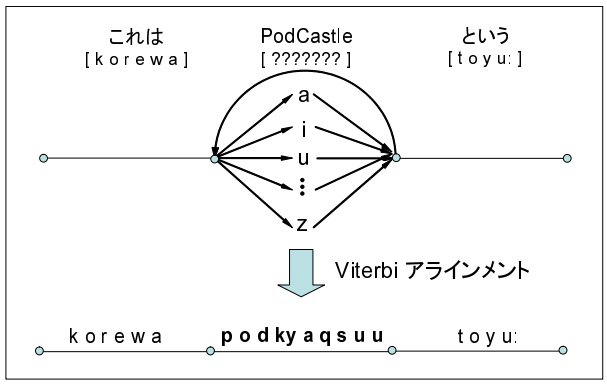

Fig. 3　未知語の音素系列の推定 (未知語「PodCastle(ポッドキャッスル)」の音素系列を推定する場合の例)

3.2.2　各単語の音素系列の獲得

ユーザからの訂正で得られるテキストを発音 (音素) 系列に変換する．ここでは，まずテキストを形態素解析し，形態素辞書に定義されている読み情報を利用して各単語ごとに音素系列を決定していく．ただし，ユーザからの訂正結果，とくにタイプ入力された区間に関しては，形態素辞書に存在しない未知語となる可能性も高く，そのような場合には正しい音素系列を求めることができない．そこで，ここでは局所的な任意音素接続ネットワークを利用することで未知語に対する音素系列の推定を行う．すなわち，図 3 に示すように，各既知語に対する正しい一意の音素系列と，未知語に対する任意音素接続のサブネットワークの両方を混在させた音素ネットワークに対して Viterbi アラインメントを行う．

3.2.3　音響モデルのパラメータ推定 (学習)

得られた学習用音素ラベルを用いて，各ポッドキャストごとに音響モデルを学習する．音響モデル学習 (話者適応) におけるパラメータの推定手法においては，一般的に，MLLR[7] などのようにモデルパラメータ間での情報の共有化を利用した方法や，MAP 推定[8] などのように学習における事前知識を効率的に利用した方法などが用いられることが多い．本研究では，これら 2 つを組み合わせた手法である MLLR-MAP を用

Table 2　テストデータ

ID	カテゴリ	エピソード数	時間 (sec.)
A	ニュース	2	2282.56
B	雑談 (独話)	2	2845.26
C	コラム	2	846.76

Table 3　音響モデル学習データ

ID	エピソード数	時間 (hour)
A	67	18.61
B	56	20.56
C	30	7.09

いた．すなわち，まず MLLR によってモデルパラメータの変換を行い，それを事前知識として MAP 推定を行う．MLLR-MAP を用いることにより，MLLR，MAP それぞれのみで学習を行う場合に比べて，高精度に学習可能であることが報告されている [9]．

4　実験と考察

実際のポッドキャストを用いた音声認識実験により，本手法の評価と考察を行った．

4.1　ポッドキャスト音声データ

本実験では 3 種類のポッドキャストをテストデータとして用いた (表 2)．A は，読み上げ音声であるが，一部の区間に背景音楽が存在する．B は，女性声優による雑談 (一般的なラジオ番組) であり，内容はエピソードごとに様々である．C は，男性タレントによるコラムであり，内容はエピソードごとに様々である．B，C のデータは，ともに自由発話音声である．

以上の 3 つのポッドキャストは，実際にユーザから比較的多くの訂正がなされている．本実験では，音響モデル学習データとして，これらの各ポッドキャストにおいて訂正が 1 カ所でもなされた全てのエピソードを用いた (表 3)．

4.2　音声認識システム

ベースラインの音響モデルは，日本語話し言葉コーパス (CSJ) から学習された triphone モデルである．このモデルは，本研究のポッドキャスト依存音響モデル学習の際の初期モデルとしても利用している．言語モデルは，Web キーワードベースの N-gram[10] であり，Web ニューステキスト，CSJ の講演書き起こしを用いて学習したものである．また，以降の認識実験ではすべて，評価用音声データの各エピソードごとに，繰り返し教師なし MLLR 適応を行っている．

Table 4　各手法での単語誤り率 (%)

ID	baseline	訂正なし	訂正あり
A	16.88	15.12	**13.24**
B	30.98	25.15	**22.21**
C	35.16	30.67	**23.54**

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 [11]．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 [6] を行い，confusion network 中の最尤候補を最終の認識結果とした．

4.3　実験結果

実験結果を表 4 に示す．ここでは比較として，事前の学習を一切行わないベースライン認識システム (「baseline」)，ユーザによる訂正結果を用いないポッドキャスト依存音響モデル学習 (「訂正なし」) の結果も併せて示す.「訂正なし」では，音声認識により自動生成した音素ラベルのみを学習時の教師信号として用いることになる．

ベースラインの結果より，認識時に入力されたエピソードごとに教師なし MLLR 適応を行っているにも関わらず，十分な認識性能は得られていない．特に，B，C の自由発話音声データに対しての性能が低く，ベースライン音響モデルと認識対象との音響的ミスマッチが，教師なし MLLR 適応だけでは十分軽減されていないことを示している．次に「訂正なし」の結果より，全てのポッドキャストに対して，ベースラインよりも一定の性能向上が得られていた．これは，同一ポッドキャストにおける音響的特性がマッチした音声データを本実験のように比較的多く用意できれば，学習時の音素ラベルが不完全であった場合でも，音響モデル学習が有効に働くことを示す.「訂正あり」の結果より，ユーザによる訂正結果を活用することで，さらに認識性能が大きく向上した．特に C のポッドキャストに対する改善が大きいが，これは C の話者が常に抑揚の大きい特徴的な発声をしており，ベースライン音響モデルとの大きなミスマッチが，本学習手法によりかなり軽減されたことが原因と考えられる．

5 おわりに

本稿では，ポッドキャスト音声認識を改善するための音響モデル学習手法について検討した．ポッドキャストのように，幅広いタスク，多様な音響的特性を持つ音声データに対し，高精度な音響モデルを学習することは従来困難であった．それに対し，本研究では，Web サービスを通じて得られる集合知 (認識誤りの訂正) の利用，ポッドキャスト依存の音響モデル学習を行うことで，性能を大きく向上させることができた．

PodCastle においては，本研究で検討したような認識誤りの訂正 (ユーザの貢献) を音声認識システムの学習に積極的に活用することは，単に認識率が向上するだけでなく，性能向上がより良いユーザ体験へと結びつき，その結果さらに多くの訂正を促すことにつながる可能性もある [2]．

本稿では，主として音響モデルに着目したが，今後は，同様の枠組みで言語モデルの学習手法，さらに両者の組み合わせによる性能評価などを行う予定である．

謝辞

本研究の一部は，科研費 (19300065) の助成を受けた．

参 考 文 献

[1] 緒方，後藤: “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] 緒方, 後藤, 江渡: “PodCastle の実現: Web2.0 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, 2007.

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

[5] 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.

[6] 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.

[7] 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.

[8] J.L.Gauvain and C.H. Lee：“Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, IEEE Trans. on Speech and Audio Processing ,Vol.2, no.2, pp.291-298(1994).

[9] E.Thelen, X.Aubert, P.Beyerlein: “Speaker Adaptation in the Philips System for Large Vocabulary Continuous Speech Recognition”, ICASSP’97, pp.1035-1038 (1997).

[10] 松原，緒方，後藤: “ポッドキャスト音声認識の性能向上手法:集合知によって更新される Web キーワードを活用した言語モデリング” 情処研報, 2008-NL-185-6, pp. 39-44, 2008.

[11] 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.