PodCastle: 集合知を活用した言語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，編集も可能な音声情報検索システム「PodCastle[1)2)3)]」の開発を行っている．ポッドキャスト音声認識では，発話内容や録音環境などが多種多様であるため，従来研究のように，タスク，ドメインに特化した大規模コーパスを事前に構築することは現実的に不可能である．それに対し，本研究ではPodCastle を通じて得られる集合知，すなわち不特定多数のユーザによる音声認識誤り訂正情報を活用した言語モデル学習を行うことで，ポッドキャスト音声認識の性能向上をはかる．

2 集合知により生成されるポッドキャストコーパス

ポッドキャストは 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．

2.1 訂正状況の分析とポッドキャストコーパス

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 [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．ここで，PodCastle の 2009 年 11 月 15 日時点における各種データ量を表 1 に示す.「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．また表 1 は，訂正された 2022 のエピソードにおける訂正単語数，全訂正のうち各訂正手段がどの程度利用されたかの内訳も示している．ボランティアベースにも関わらず，この時点までに 44 万単語もの訂正が得られている．訂正手段の内訳としては，候補選択による訂正がより多く利用されていた．実際の利用状況を確認すると，1-best の認識結果に誤りが多くても，競合候補中に本来の正解がある程度含まれていると，候補選択による訂正インタフェースが積極的に利用される傾向にあった．ただし，現状ではそもそも競合候補にも正解がほとんど含まれないような認識困難なポッドキャストも多く存在し (例えば芸能人の会話番組など)，そのようなデータではタイプ入力が主な訂正手段となっている．

Fig. 1 音声訂正インタフェース

Table 1 PodCastle における各データ量

登録済みポッドキャスト数	621
エピソード (mp3 音声ファイル)	63278
訂正されたエピソード数	2022
総訂正単語数	440570
平均訂正単語数/エピソード	218
訂正手段 (候補選択)	227293
訂正手段 (タイプ入力)	213277

このように，実際にユーザが訂正することで生成された書き起こし，並びにそれらに対応する音声データにより**ポッドキャストコーパス**が構成される．前述したようにポッドキャストは多種多様な音声データであるため，本コーパスはタスクに非依存な特性を持っている．また最も重要なポイントとして，Web サービスの利用に伴って音声，テキストデータともに日々蓄積され，コーパスの規模が徐々に拡大していくことが挙げられる．

3 集合知を活用した言語モデル学習

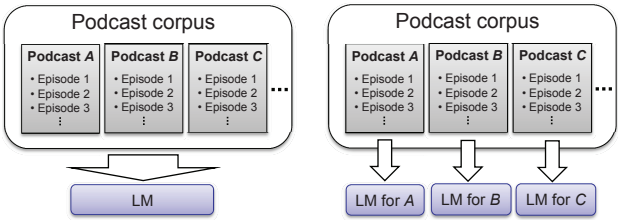
ここでは，ポッドキャスト音声認識における言語モデルの**汎用性**と**適応性**の観点から以下で述べる 2 つの学習アプローチを導入する．

3.1 ポッドキャスト非依存言語モデリング

PodCastle においては，様々なタスク，ドメインの書き起こしデータが収集される．本アプロー

---

*PodCastle: Collaborative Training of Language Models Based on Wisdom of Crowds for Automatic Speech Recognition
by Jun Ogata, Masataka Goto (AIST)



(a) *Podcast independent LM* (b) *Podcast dependent LM*

Fig. 2　2 つの言語モデル学習アプローチ (LM: Language Model)．実際は，各言語モデルは他タスク言語モデルと混合される．

チは収集されたポッドキャストコーパス全体を利用することで，様々なポッドキャストに対して一定の性能が得られる，汎用的な言語モデルの構築を目指すものである (図 2-(a))．なお，一般的に音声認識のための統計的言語モデルにおいては，幅広いタスクの膨大なテキストを学習に利用するよりは，小/中規模でもよいので該当するタスクに特化したテキストのみで学習する方が性能/効率ともによいとされる [4]．ただしポッドキャスト音声認識では，事前に対象音声のトピックや発話スタイル，語彙などの特定ができないため，(少なくとも初期段階の認識においては) このような汎用性を目指すアプローチが有効に働くと考えられる．また，後述するポッドキャスト依存言語モデリングにおける初期モデルとしても有効となる．システムの概要を以下に示す．

まず，事前に複数のコーパスを用いてそれぞれのタスク，ドメインに依存した言語モデル (要素モデルと呼ぶ) を構築する．本研究では，日本語話し言葉コーパス (CSJ)，Web ニュースコーパス，そして前節で述べたポッドキャストコーパスを利用することで 3 種類の要素モデルを学習する．CSJ では主に話し言葉への対応，Web ニュースコーパスでは幅広いトピック，最新の単語への対応をそれぞれねらっている．一方，ポッドキャストコーパスにおいては，様々なタスクにおける音声データの書き起こしが得られるため，幅広いトピック，講演口調以外の様々な発話スタイルや会話等への対応などが期待できる．

次に，得られた要素モデルをモデルベースで混合することで，最終的な言語モデルを構築する．ここで混合手法としては線形補間法を用いる．線形補間法では，補間パラメータ (混合重み) をいかに最適に設定するかが重要となる．本研究では，現段階までに蓄積されたポッドキャストコーパスの汎用性を評価するために，様々なトピック，発話スタイルのポッドキャストから構成される開発セットを補間パラメータの最適化に用いる．最適化には，開発セットのパープレキシティが最小となるように，EM アルゴリズムによる繰り返し推定を行う．

3.2　ポッドキャスト依存言語モデリング

ポッドキャスト音声認識の更なる性能向上につなげるために，ポッドキャストにおける個々のタスク，ドメインに特化した (適応された) 言語モデルの構築を目指す．本研究では，その 1 つとして PodCastle に登録されている個々のポッドキャストごとに言語モデルの学習を行う，ポッドキャスト依存言語モデリングを導入する (図 2-(b))．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，同じ言語的特性 (トピック，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言語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．ポッドキャスト依存言語モデリングの流れを以下に示す．

1. あるポッドキャスト A において，訂正がなされたエピソードの全書き起こしを利用して，ポッドキャスト A の要素モデルを学習する．
2. ポッドキャスト A の要素モデルとベースラインモデル (ここでは CSJ と Web ニュースを混合したモデル) を線形補間する．この段階では初期パラメータとして $\lambda = 0.5$ と設定する．
3. 対象となる音声データ (新たなエピソード) に対して，(2) で構築した言語モデルにより音声認識を行い，認識仮説を生成する．
4. 認識仮説を開発セットとして，EM アルゴリズムにより補間パラメータの最適化を行う (対象音声データに特化した補間パラメータ)．
5. 最適化された補間パラメータにより再度線形補間を行い，ポッドキャスト A 依存の言語モデルを構築する．

4　実験

4.1　ポッドキャスト非依存言語モデリングの評価

本実験で利用する各学習セット，評価セット，開発セットについて表 2 にまとめる．評価セットは 7 つのポッドキャストで構成されており，主に，ニュース，経済コラム，対談，雑談，4 種類に内容的に分類できる．経済コラムは，発話スタイルとしては学会講演に近いが，雑音 (背景音楽)

Table 2　学習，評価セット各データ量

	文数	単語数
CSJ	389,309	7,043,529
Webnews	18,190,294	456,017,101
PodcastALL	412,047	4,350,825
PodcastUSER	83,672	1,467,596
評価セット	2,088	49,634
開発セット	1,728	35,346

Table 3　各言語モデルの評価

	PP	APP(#OOVs)	WER
CSJ	153.94	245.46(3455)	52.22%
Webnews	548.36	583.80(594)	58.20%
CSJ+Webnews	180.49	189.23(459)	47.86%
+PodcastHypo	155.82	161.65(371)	47.04%
+PodcastUSER	155.74	161.81(383)	**46.63%**
+PodcastALL	151.11	**156.33(348)**	46.81%

を含んでいる．また，対談，雑談については日常会話と同等のカジュアルな発話スタイルであり，発話速度も速いため，非常に認識が困難なタスクとなっている．開発セットは，同じ 7 つのポッドキャストのうち，評価セットに含まれないエピソードで構成される．

本実験で構築した全ての言語モデルは 3-gram である．CSJ は，日本語話し言葉コーパス (CSJ) の講演書き起こし，Webnews は 2006 年 8 月〜2009 年 11 月における Yahoo!ニュースの記事である．PodcastALL はポッドキャストコーパス中のデータであり，評価セットのポッドキャストは含んでいない．PodcastALL はユーザによる訂正箇所以外は全て音声認識結果となっている．ここでは比較として，ユーザによる訂正がなされた発話 (文) のみを学習に利用する学習セット PodcastUSER も考慮する．各モデルの評価には，評価セットに対するパープレキシティ (PP) と補正パープレキシティ (APP)，単語誤り率 (WER) を用いた．補正パープレキシティは，評価データ中の未知語 (OOV) 出現率を考慮したパープレキシティであり，下記の式で表される [6]．

$$APP = (P(w_1 \dots w_n) m^{-n_u})^{-\frac{1}{n}} \qquad (1)$$

ここで n は評価セットの総単語数，n_u,m は評価セット中の未知語の総数，種類数をそれぞれ示す．本実験では，音声認識システムとしてはシンプルなワンパスのデコーディングを用いており，音響モデルの教師なし適応等による多段処理は行っていない．

4.1.1　実験結果

各言語モデルの評価を表 3 に示す．まず上段の CSJ，Webnews の要素言語モデルについて考察する．CSJ，Webnews を単独で用いた場合，PP，OOV に関してそれぞれ一長一短があることがわかる．Webnews は話し言葉でなく基本的に書き言葉であるため，パープレキシティ自体は大きくなるが未知語を劇的に減らすことができている．これら 2 つを線形補間したモデル CSJ+Webnews は，それぞれの特徴を表現したモデルとなっており，APP，WER ともに大きく削減できていた．

表中，下段 3 つは CSJ+Webnews とそれぞれ線形補間したモデルの結果である．PodcastHypo は，ユーザの訂正結果を含まず，全て認識結果を学習テキストとしたモデルである．ただし，ここでの認識結果は，PodCastle の実際の運用上で得られた認識結果であり，本実験と同一の認識システム，モデルを利用して得られたものではない．ポッドキャストコーパスの要素モデルを線形補間することで，いずれの場合においても CSJ+Webnews と比較して性能が改善していることがわかる．評価セット中の全てのポッドキャストに対して改善が得られ，特に対談や雑談など，CSJ+Webnews ではカバーできない発話スタイルや言い回しに対する改善が大きかった．パープレキシティ，補正パープレキシティについては PodcastALL が最もよい性能を示したものの，WER についてはユーザからの訂正を含む発話のみを用いた PodcastUSER が最も削減率が大きい結果となった．これは PodcastALL では，PodcastUSER と比較して，単純にデータ量が多いため PP，APP ではより良い値を示したものの，誤認識単語が言語モデル学習の際に考慮されるため，その結果 WER が高くなったと考えられる．これに対しては今後，認識仮説の信頼度等の基準による発話選択処理を導入することにより更なる向上が見込める．

4.2　ポッドキャスト依存言語モデリングの評価

次に，ポッドキャスト依存言語モデリングの有効性を音声認識実験により評価する．本実験では 3 種類のポッドキャストを評価セットとして用いた (表 4)．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，男性芸能人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

Table 4　評価セット

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 5　学習セット

ID	エピソード数	単語数
A	67	270,447
B	56	283,414
C	30	39,098

Table 6　各手法での WER(%)

ID	ベースライン	訂正なし	訂正あり
A	16.88%	16.24%	**14.49%**
B	30.98%	28.52%	**24.61%**
C	35.16%	33.22%	**26.24%**

る．本実験では，言語モデル学習セットとして，これらの各ポッドキャストにおいて訂正が 1 箇所でもなされた全てのエピソードを用いた (表 5)．ただし，評価セットのエピソードは学習セットには含んでいない．音声認識には PodCastle 音声認識システムを用いた [2]．ベースラインとなる言語モデルは，Web キーワードベースの N-gram[7] 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，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．

4.2.1　実験結果

実験結果を表 6 に示す．ここでは比較として，ポッドキャスト依存言語モデル学習を行わないベースライン認識システム (「ベースライン」)，ユーザによる訂正結果を用いないポッドキャスト依存言語モデル学習 (「訂正なし」) の結果も併せて示す.「訂正なし」では，音声認識結果のテキストのみを用いて言語モデル学習を行う．実験結果より，ポッドキャスト依存言語モデリングによっていずれのポッドキャストに対しても性能改善が得られていることがわかる．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 の話者が独特の発話スタイル，言い回しを持っているためであった (C の番組内容はクイズ形式のコラムであり，他のポッドキャストと比べても異質な内容であった)．B に関しては，内容はエピソードにおいて様々であるが，独自の言い回しや冒頭，終了部分での特定のフレーズ，曲名やアルバム名等の専門用語に対して特に改善がみられた．なお本実験は，4.1 節のポッドキャスト非依存言語モデリングとは独立した実験として評価を行ったが，これら 2 つのアプローチは併用して更なる改善を得ることも可能である．

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言語モデル学習手法について検討した．我々は以前の報告で，4.2 節の評価実験と同等の条件にて音響モデル学習の性能改善に関する実験を行った (集合知に基づくポッドキャスト依存音響モデル学習)[3]．今回の実験は全体の傾向として，音響モデル学習の性能改善に比べて数値的には低い結果となっている．ただし，言語的要因による改善では，個々のポッドキャスト独特のトピックやキーワード，前述したような専門用語など音声データ中の重要箇所に直接影響するため，改善による意義は大きいと考えられる．また，PodCastle においては，ユーザがあるエピソードの訂正時に入力した特徴的な単語が，本研究の言語モデル学習によって別エピソードで正しく認識できるようになることで，訂正してくれたユーザに対してより良い印象を与えることができる．加えて，それによってユーザ自身がシステムへの貢献を実感することで，さらなる訂正の促進につながる可能性もある．

今後は，発話選択処理の導入を行うとともに，音響モデル学習との併用によるさらなる性能改善についても調査していく．

参 考 文 献

[1] 後藤，緒方，江渡: “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”, 人工知能学会論文誌, Vol.25, No.1, pp.104–113 (2010).

[2] J.Ogata, M.Goto, K.Eto: “Automatic Transcription for a Web 2.0 Service to Search Podcasts”, Interspeech 2007, pp.2617–2620 (2007).

[3] J.Ogata, M.Goto: “PodCastle: Collaborative Training of Acoustic Models on the Basis of Wisdom of Crowds for Podcast Transcription”, Interspeech 2009, pp.1491–1494 (2009).

[4] 伊藤: “音声認識における言語モデル”, 音響誌, Vol.66, No.1, pp.32–35 (2010).

[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] 中川, 赤松: “未知語を含む文集合のパープレキシティの算出法-新補正パープレキシティ-”, 音講論集, pp.63–64 (1998).

[7] 緒方, 松原, 後藤: “PodCastle: 集合知に基づく Web キーワードを考慮した言語モデリング”, 音講論集, pp.97–100 (2008).