PodCastle: 動的トピック混合に基づく教師なし言語モデル適応 *

○緒方 淳，後藤 真孝 (産総研)

1 はじめに

我々は Web 上の代表的な音声コンテンツの 1 つである**ポッドキャスト**を対象とした音声情報検索 Web サービス「**PodCastle**[1)2)3)]」の開発・運用を行っている．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，従来の大語彙連続音声認識システム [4)5)6)] 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．したがって，事前コーパスに依存することなく，いかに高精度な音響モデル，言語モデルを構築，学習するかが性能向上への鍵となる．特に言語モデル (N-gram) は単純なモデル構造であるが故に，音響モデルに比べ，学習データにより強く依存する傾向があり [7)]，ポッドキャスト音声認識性能を劣化させる大きな要因となっている．

本研究では，Yahoo!ニュース，Google ニュース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**) を有効活用し，ポッドキャストの多様なトピックに頑健な言語モデルを構築することを目指す．

2 動的言語モデリング

言語モデル適応については，これまでにも幾つかのタスク (放送ニュース [4)8)]，ミーティング [9)10)]，講義 [11)12)]) に対して，様々な研究がなされている．これらの研究では基本的に，各タスクにマッチした大量のテキストデータから学習したメイン言語モデルに対して，ドメイン内 (トピック一致) テキスト [8)9)]，Web ベーステキスト [4)12)13)]，ユーザによるフィードバックや書き起こし [10)11)] といった認識対象に関連する付加的なテキストデータを用いて適応する．しかし，本研究で対象とするポッドキャストにおいては，このようなタスクに合致したメイン言語モデル自体を用意することはできない．そこで，本手法では，様々なトピックをカバーする大量の Web ニューステキストをベースにしてメイン言語モデルを構築し，さらにその特性を活かして，認識対象ごとのトピックに合致するよう言語モデルのパラメータを最適化することで動的な言語モデル適応 (**動的言語モデリング**) を行う．

2.1 Web ニューステキスト

Web ニューステキストには，音声認識の言語モデリングにおいて有用となり得る 2 つの大きな特徴がある．まず 1 つ目は，一般的なニュースアグリゲーション Web サイトでは，様々なニュース配信サービスからの幅広い内容に関するニュース記事が集約されており，それらの記事はユーザが閲覧しやすいように複数のトピック，カテゴリごとに分類されている．そして 2 つ目は，日常的に記事が更新される仕組みにより，一般社会における最新のトピック・語彙がカバーされている点である．本研究では，Web ニューステキストとして Yahoo! Japan ニュース (http://headlines.yahoo.co.jp/hl) の膨大な記事データを利用する．Yahoo! Japan ニュースでは，全てのニュース記事が 6 メイントピック，25 サブトピックからなる階層構造上に分類されている．ここでは，2007 年 2 月〜2010 年 6 月の 40ヶ月間に配信されたニュース記事を言語モデリングに利用する．表 1 に，各トピック，サブトピックにおける名称とデータ量を示す．本研究では，表中のサブトピックごとに要素言語モデルを構築するため，以降では便宜上，この 25 のサブトピックを単に「トピック」と示すことにする．

2.2 トピック言語モデルの動的混合

Web ニューステキストに基づくトピック言語モデルを利用して，適応言語モデルを動的に生成する．本研究で構築した動的言語モデリング手法を図 1 に示す．本システムは，各トピック言語モデルを用いたモデルレベル混合手法 [14)] に基づいている．モデルレベル混合では，複数の要素モデルの N-gram 確率を下記のように重み付きで補間する．

$$p_{mix}(w|h) = \sum_i \lambda_i p_i(w|h) \qquad (1)$$

ここで λ_i は，$\sum_i \lambda_i = 1$ を満たす混合パラメータ (重み) である．一般的に，各要素モデルの混合重みは，評価セットと同一タスクのヘルドアウトセットを用いて最適化する．最適化手法としては，ヘルドアウト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が用いられる．

本システムにおける静的プロセスとして，まず Web ニューステキストから表 1 に示す 25 分野のトピック言語モデルを学習する．ここで，ポッドキャスト音声中の話し言葉口調に対処するために，別の要素モデルとして日本語話し言葉コーパス (CSJ)[5)] から学習した言語モデルを用意し，それぞれのトピック言語モデルと線形補間を行う．この際の補間重みは 0.5 とした．また，各トピック言語モデルの語彙は，各トピッ

* PodCastle: Unsupervised Language Model Adaptation Based on Dynamic Topic Mixture
by Jun Ogata, Masataka Goto (AIST)

Table 1　Web ニューステキストデータ量 (2007 年 2 月〜2010 年 6 月に配信されたニュース記事)

トピック (サブトピック)	単語数	トピック (サブトピック)	単語数
経済 (市況)	8.3M	エンターテインメント (その他)	43.1M
経済 (株式)	10.4M	スポーツ (野球)	23.4M
経済 (産業)	23.5M	スポーツ (サッカー)	14.3M
経済 (その他)	55.7M	スポーツ (モータースポーツ)	5.6M
国内 (政治)	19.3M	スポーツ (競馬)	5.9M
国内 (社会)	65.3M	スポーツ (ゴルフ)	7.4M
国内 (人)	0.7M	スポーツ (格闘技)	8.8M
海外 (中国)	16.6M	スポーツ (その他)	50.1M
海外 (韓国)	9.0M	テクノロジー (インターネット)	7.3M
海外 (その他)	32.7M	テクノロジー (モバイル)	5.9M
エンターテインメント (音楽)	14.0M	テクノロジー (セキュリティ)	2.2M
エンターテインメント (映画)	10.6M	テクノロジー (その他)	48.7M
エンターテインメント (ゲーム)	9.3M		

クテキストから頻度順で選択した 60000 単語と CSJ テキスト中の語彙 20000 単語をマージしたものを用いた．次に，これら 25 のトピック言語モデルをモデルレベルで混合することで，全てのトピックの要素を表現する単一の初期言語モデルを生成する．一般的に，Web ニューステキストは様々なトピックをカバーするが，表 1 の例にも見られるようにトピックごとのデータ量にある程度の偏りがある．ここでの初期言語モデルは，ポッドキャスト中の様々なトピックに対して一定の性能を得ることのできるグローバルなモデルとするために，各トピックモデルを同一の重み ($\lambda_i = 1/25$) でモデル混合を行う．初期言語モデルの語彙サイズは，25 の各トピック言語モデルの語彙 (約 60000 単語) を全てマージした 286345 単語とした．

入力音声 (ポッドキャストエピソード) ごとの動的プロセスとして，まず，上記初期言語モデルを用いて音声認識を行い，初期認識結果を生成する．そして，初期認識結果を用いて各トピック言語モデルの混合重みを動的に算出する．すなわち，初期認識結果のテキストを前述のヘルドアウトセットとして，混合結果のモデルが最小のパープレキシティを示すように EM アルゴリズムにより混合重みを推定する．そして，最適化した混合重みを基にトピック言語モデルを混合し，入力音声のトピックに適応化した最終的な言語モデルを出力する．

2.3　ポッドキャスト依存言語モデルの利用

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，ポッドキャストごとのトピック，ドメインに特化した言語モデルの構築を目指す．ここでは，認識対象エピソードと同じポッドキャスト内の他の (過去の) エピソードデータを利用して言語モデルを構築し (**ポッドキャスト依存言語モデル**)，これを前述の動的言語モデリングに組み込む．この理由としては，同一の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．さらに，ポッドキャストを構成する RSS の仕組みにより，認識対象となる各エピソード音声データがどのポッドキャストに属するのか，すなわち，各音声ごとにどの言語モデルを教師なし適応時に適用すべきかが自明であるという利点もある．

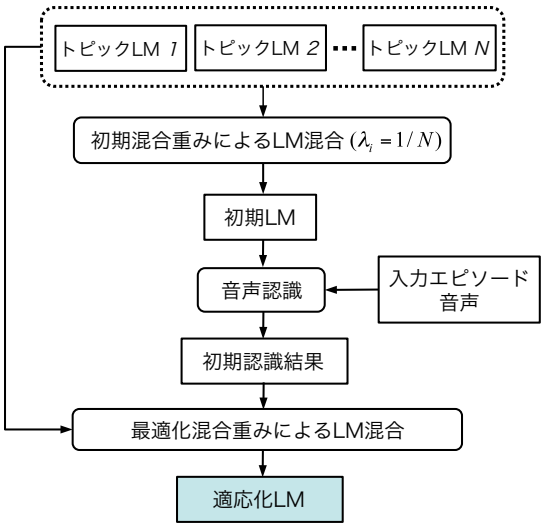

Fig. 1　動的言語モデリング手法 (カジュアルな発話スタイルに対処するために，各トピック言語モデルは事前に話し言葉テキスト (CSJ) と線形補間を行っている．)

拡張システムでは，まず事前にポッドキャスト依存言語モデルを，認識対象エピソード以外の過去のエピソードを利用して学習しておく．この際，我々の PodCastle システムでは，過去のエピソードのテキストデータとしてユーザ貢献により訂正された書き起こしを利用することも可能であるが [15]，本研究では主として教師なしアプローチによる動的言語モデリングを検討するために，音声認識により自動的に書き起こされたテキストを用いる．そして，ポッドキャスト依存言語モデルは，図 1 に示す最終的なモデルレベル混合処理において付加的な要素モデルとして

追加する．トピック言語モデルとともに混合重みを前述の手法で自動推定し，最終的な適応化言語モデルとして混合する．

3 実験

前章で述べた動的言語モデリング手法に対して，実際のポッドキャスト音声データを用いて評価実験を行う．

利用するポッドキャスト音声データの諸元について表 2 にまとめる．ここで，評価セットは実際に音声認識性能を評価するためのデータであり，学習セットは各ポッドキャスト依存言語モデルの学習に利用したデータである．評価セットは 8 ポッドキャスト，合計 47 エピソードで構成されており，ドメインとしてはデイリーニュース，政治・経済のコラム，レクチャー形式のトーク，雑談に大きく分類できる．トピックについてもポッドキャストごとに様々であり，ニュース番組 (A,B) においては 1 つのエピソード内でもスポーツ，政治，経済といった複数のトピックが存在する．

音声認識には，PodCastle 音声認識システムを用いた [16]．音響モデルは，CSJ の約 600 時間の講演音声データから学習された，状態数 3000，1 状態あたり混合ガウス分布数 16 の tied-state cross-word triphone モデルである．特徴量には 39 次元 PLP(12 次元 PLP ケプストラム係数と正規化パワー，それらの Δ，ΔΔ)，そして話者，環境の変動に対処するために CMLLR ベースの適応化学習 [17] を行っている．

3.1 実験結果

表 3 に本研究で構築した動的言語モデリング手法の認識性能を示す．表中，ベースラインは，動的言語モデリングにおける初期言語モデル (初期混合重みでトピック言語モデルを混合したモデル，図 1 中の “初期 LM”) を用いた際の認識性能である．また，音声認識結果テキストを利用した混合重み自動推定における認識誤りの影響を調査するために，混合重み自動推定に正解書き起こしを用いた教師あり実験も行った．構築した認識システムは，教師なし音響モデル適応を含めたマルチパスデコーディングに基づくが，本研究では言語モデルにおける純粋な比較評価を行うため，各実験において共通の音響モデル (図 1 の “初期認識結果” で MLLR 適応した音響モデル) を用いた．

3.1.1 動的言語モデリングの性能評価

まず，ポッドキャスト依存言語モデルなしのシステムの結果 (*podcast LM* なし) より，評価セット中の全てのポッドキャストにおいてベースラインに比べての改善がみられた (教師ありの場合に絶対値で 1.3%，教師なしの場合に 1.2%の改善)．本手法での混合重み最適化手法は，ポッドキャストエピソードごとにパープレキシティ最小化基準で可能性のあるトピックを複数選択することに相当する．特に大きな改善が得られたポッドキャスト (B, G) では，本最適化手法によって内容に合致したトピック (B の場合は複数) が選択され，λ_i の値も全 25 トピックの中で支配的であった．混合重み自動推定における教師ありと教師なしの比較では，最終的な単語誤り率は両者において大きな差はなく，絶対値で 0.1%程度であった．このような傾向は文献 9) のミーティングタスクにおいても示唆されており，混合重み自動推定は音声認識誤りにある程度頑健であるといえる．

3.1.2 ポッドキャスト依存言語モデルの効果

最後に，ポッドキャスト依存言語モデルを利用した動的言語モデリングの性能について述べる．表 3 の “*podcast LM* なし” より，認識性能がさらに改善され，教師なしの場合で最終的に 32.4%の単語誤り率を得た (ベースラインと比べて絶対値で 2.5%の改善)．ここでの傾向としては，ポッドキャスト依存言語モデルの学習データが多いポッドキャストほど，より大きな性能改善が得られている．これにより，学習テキストが誤りを含む音声認識結果であっても，ポッドキャストの単位で学習することで言語モデルにおけるトピックをある程度表現することができるといえる．また，ここでの混合重み自動推定においても教師ありと教師なしとで大きな差はなかった．ポッドキャスト依存言語モデルを用いた動的言語モデリング手法は，全ての処理が教師なしで実行されるため，ポッドキャスト音声認識，そして PodCastle Web サービス運用において有用だといえる．PodCastle ではさらに，ユーザ貢献により訂正された書き起こしを学習に利用することができ，本研究で構築した動的言語モデリングをより効果的に行うことも可能になる．

4 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リングについて検討した．ポッドキャストのように，幅広いタスク，多様な言語的特性を持つ音声データに対し，高精度な言語モデルを学習することは従来困難であった．それに対し，本手法では Web ニューステキストを有効活用することで，入力エピソードごとに，動的なトピック混合に基づき言語モデルを教師なし適応していく．実際の日本語ポッドキャスト音声データにより評価を行ったところ，Web ニュースベースのトピック言語モデルのみを用いた動的適応で 3.4%の改善率が得られ，さらにポッドキャスト依存言語モデルを考慮することで 7.2% の改善が得られた．

本研究で着目した Web ニュースは，一般社会にお

Table 2　ポッドキャスト音声データ (学習セットの単語数は音声認識結果の単語数を示す)

ID	ドメイン	トピック	評価セット エピソード数 (単語数)	学習セット エピソード数 (単語数)
A	ニュース	複数	4 (11170)	383 (1027390)
B	ニュース	複数	4 (4937)	496 (985273)
C	コラム	政治	20 (13876)	2189 (1591478)
D	コラム	経済	5(10763)	215 (743858)
E	レクチャー	株式	6(5315)	52 (54468)
F	レクチャー	ヘルスケア	2(3292)	119 (259457)
G	雑談	野球	2(4439)	15 (37874)
H	雑談	芸能	4(14590)	98 (458936)

Table 3　動的言語モデリング手法の認識性能 (単語誤り率 (%)). *podcast LM* はポッドキャスト依存言語モデルを示す．“教師あり” は混合重み最適化に正解の書き起こしを利用した場合，“教師なし” は音声認識結果を利用した場合を示す．

ID	ベースライン	動的言語モデリング			
		podcast LM なし		*podcast LM* あり	
		教師あり	教師なし	教師あり	教師なし
A	17.9	16.2	16.4	14.0	14.2
B	21.3	19.2	19.3	17.4	17.3
C	28.2	27.4	27.2	26.3	26.3
D	41.1	39.6	39.8	38.1	38.3
E	18.8	17.0	17.0	16.2	16.6
F	29.7	28.8	28.2	25.1	25.1
G	51.0	49.0	49.0	48.9	48.7
H	56.7	55.6	56.2	54.9	55.1
Ave.	34.9	33.6	**33.7**	32.2	**32.4**

いて関心の高い様々な最新のトピックを総合的に集約したものであるといえる．したがって，音声認識の言語モデルとしては，ポッドキャストだけでなく様々なタスク，ドメインにおいて有効に働く，汎用性の高いモデルとなっていると考えられる．今後は，ポッドキャスト以外の様々なデータに対して動的言語モデリング手法の効果を検証していく．また，動的言語モデリングの性能を改善させるために，より高度な言語モデル補間手法，未知語を考慮した語彙選択手法なども検討する予定である．

参 考 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] J. Ogata and M. Goto: PodCastle: Collaborative Training of Acoustic Models on the Basis of Wisdom of Crowds for Podcast Transcription, *Proc. of Interspeech 2009*, pp. 1491–1494 (2009).

[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，人工知能学会論文誌，**25**，104–113 (2010).

[4] M. Federico and N. Bertoldi: Broadcast news LM adaptation over time, *Computer Speech & Language*, **18**，417–435 (2004).

[5] T. Kawahara, H. Nanjo, T. Shinozaki and S. Furui: Benchmark test for speech recognition using the corpus of spontaneous japanese, *Proc. SSPR 2003* (2003).

[6] Y. Akita, M. Mimura and T. Kawahara: Automatic Transcription System for Meetings of the Japanese National Congress, *Proc. of Interspeech 2009* (2009).

[7] F. Lefevre, J.-L. Gauvain and L. F. Lamel: Genericity and portability for task-independent speech recognition, *Computer Speech & Language*, Vol. 19, pp. 345–363 (2005).

[8] X. Lei, W. Wu, W. Wang, A. Mandal and A. Stolcke: Development of the 2008 SRI mandarin speech-to-text system for broadcast news and conversation, *Proc. of Interspeech 2009* (2009).

[9] G. Tur and A. Stolcke: Unsupervised language model adaptation for meeting recognition, *Proc. ICASSP2007* (2007).

[10] D. Vergyri, A. Stolcke and G. Tur: Exploiting user feedback for language model adaptation in meeting recognition, *Proc. of ICASSP 2009* (2009).

[11] B.-J. P. Hsu and J. Glass: Language model parameter estimation using user transcription, *Proc. of ICASSP 2009* (2009).

[12] 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).

[13] 増村　亮，伊藤　仁，伊藤彰則，牧野正三：WWW を利用したトピック関連語推定に基づく言語モデル教師なし適応の性能評価，情処研報 音声言語情報処理 2010-SLP-79-33 (2009).

[14] F. Jelinek and R. L. Mercer: Interplated estimation of Markov source parameters from sparce data, *Proc. Workshop on Pattern Recognition in Practice* (1980).

[15] 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 2009-SLP-80-10 (2009).

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

[17] M. J. F. Gales: Maximal likelihood linear transformations for HMM-Based speech recognition, *Computer Speech & Language*, **12**，75–98 (1998).