Vol.2010-SLP-84 No.2
2010/12/20

PodCastle: 動的言語モデリングに基づくポッドキャスト音声認識

緒方　淳[†1]　後藤真孝[†1]

ポッドキャスト音声認識では，トピック，語彙，発話スタイルといった言語的特性の変動が大きいため，言語モデルとしてそれらをいかにモデル化するかが重要な課題となる．本稿では，ポッドキャスト音声認識の性能向上のための動的言語モデリング手法について述べる．本研究のアプローチでは，多様なトピックをカバーする Web ニューステキストデータを利用し，トピックごとに学習された複数の言語モデルを動的に混合していく．この際，言語モデルパラメータ (混合重み) は，入力音声データごとに教師なしで自動調整する．さらに，実際のポッドキャストデータから学習された言語モデルを動的言語モデリングの要素モデルとして考慮する．動的言語モデリング手法を実際の日本語ポッドキャスト音声データに対して音声認識実験を行ったところ，従来の言語モデルと比べて性能改善が得られた．

PodCastle: Dynamic Language Modeling for Podcast Transcription

Jun Ogata[†1] and Masataka Goto[†1]

In transcribing podcasts, language modeling is a critical and challenging issue because of the diversity of topics, vocabularies, and speaking styles. This paper describes the application of a dynamic language model adaptation to podcast transcription. Our approach is based on a dynamic mixture of multiple topic language models through the use of web news text data. In the mixture process, the topic language model parameters are automatically tuned in an unsupervised manner for each input speech data. Furthermore, we explore the use of podcast-dependent language models trained from actual podcast data for the dynamic language modeling process. Experiments have been carried out for Japanese podcast transcription. We confirmed that the adapted language models in our system reduced the word error rate significantly compared to the baseline language model.

1. はじめに

YouTube に代表される動画共有サービス，ポッドキャストの普及により，Web 上では大量の音声コンテンツが日常的に生成・蓄積されるようになった．特に最近は，米国大統領選挙，国内では政府による行政刷新会議など，社会的関心の高い出来事はこうしたコンテンツとして積極的に公開・利用され，政治・経済・社会の様々な場面で大きい影響を与えるようになっている．このような Web 上の膨大なコンテンツからユーザが欲しい情報を自由に検索したり，快適な鑑賞を行うためには，コンテンツに含まれる音声情報を計算機が的確に理解し，索引情報を付与する「**Web 音声インデキシング**」が重要となる．

一方，最近では，音声認識技術の高度化ならびに実環境音声データを対象とした大規模なコーパスが整備され始めたことにより，アーカイビング，情報検索，要約，字幕化，翻訳など様々な**音声ドキュメント処理**に関する研究が精力的に進められるようになった．しかしながら現状では，Web 音声インデキシングをはじめとする，Web 上の膨大かつ多様な音声コンテンツを対象とした音声ドキュメント処理は，基盤技術となる音声認識が困難であるため，Google，Yahoo!等に代表されるテキスト検索のような，日常的に利用されるシステムは実現されていない．

このような状況の中，我々は Web 音声インデキシングの一環として，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)]」の開発を行っている．PodCaslte は，ポッドキャストを音声認識技術によって自動的にテキスト化・索引付けすることで，それらをユーザが全文検索できるだけではなく，Web ブラウザを通じて詳細な閲覧，編集も可能にする「ソーシャルアノテーションシステム」である．本研究では，そのようなシステムを実現・運用し，ひいては高度な Web 音声インデキシングを実現するために，Web 上の実環境音声コンテンツであるポッドキャストを対象とした音声認識技術について検討を行っている．

従来，実環境音声の認識に関する研究においては，放送ニュース[4)]，学会講演[5)]，国会中継[6)] 等を対象とした報告がなされており，いずれの場合においても，それぞれのタスク，ドメインにマッチしたコーパスを事前に用意し，音声認識器 (音響モデル，言語モデル) を学習することで大きな改善が得られている．一方，本研究で対象とするポッドキャストは，そ

†1 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

の発話内容や録音環境などが多種多様であるという特徴を持っている．そのため，ポッドキャスト上で出現する全てのタスク，ドメインに対して，従来研究のように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，ポッドキャスト音声認識においては，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．

そこで本稿では，ポッドキャスト音声認識のための言語モデリング手法について検討する．我々はこれまでに，ポッドキャストのような多様な音声データのための言語資源として，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**) に着目し，言語モデル学習データとして利用することの有効性を確認している[8)9)]．本研究では，Web ニューステキストをより有効活用し，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．具体的には，Web ニュース上の幅広い複数のトピックに分類されたテキスト記事をもとに，各トピックに依存した要素モデル (**トピック言語モデル**) 生成し，認識対象の入力音声データごとに適切なトピックが反映された適応言語モデルを動的に構築していく．

2. 動的言語モデリング

ポッドキャスト音声認識のための，動的な言語モデル適応手法について述べる．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)10)]，ミーティング[11)12)]，講義[13)14)]) 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[10)11)]，Web ベーステキスト[4)14)]，ユーザによるフィードバックや書き起こし[12)13)] 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といえる．まず，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 つ目としては，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．

表 1 トピックごとの 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		

本研究では，言語モデルにおけるトピックの多様性に対処するために，ニュースアグリゲーション Web サイトの 1 つである Yahoo! Japan ニュース[⋆1]の膨大なニュース記事を利用する．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 Web キーワードを利用した形態素解析

Web ニュースは日々の最新の単語，専門用語を多く含むため，一般的な形態素解析器 (形態素解析辞書) では，言語モデル学習の事前処理として行われる単語分割において分割誤りが顕著となる (例：“ケータイ” ⇒ “ケー”, “タイ”)．そこで，不特定多数のユーザによって

⋆1 http://headlines.yahoo.co.jp/hl

日々整備，更新されている「Web キーワード辞書」を活用した形態素解析を行うことで，そのような新出語の分割誤りを低減し，さらに読み (発音) の情報も獲得する[15]．本研究では Web キーワード辞書として「はてなダイアリーキーワード⋆1」を利用する．

2.3 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システムを図 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[16] 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．

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

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にあらわれる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，算出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．本研究では，最終的なモデル混合手法として以下の 2 種類を検討する．

- **全モデル混合**:
  全てのトピック言語モデルを，前述の自動推定した重みをもとに混合する．
- **選択的モデル混合**:
  前述の自動推定した重みの値が一定以上のトピック言語モデルのみを選択し混合を行う．すなわち，初期認識結果に対してパープレキシティが低くなる上位いくつかのトピック言語モデルのみを用いる．最終的なモデル混合の際には，選択されたトピック言語モデルのみを用いた場合の混合重みを再推定する．

前者の全モデル混合では，最終言語モデルの語彙は初期言語モデルと同一の 286345 単語となる．一方，後者の選択的モデル混合では，最終言語モデルの語彙は，入力されたエピソードごとに選択されたトピック言語モデルの語彙サイズに絞り込むことができる．

2.4 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システムに組み込む．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動的言語モデリング時に適用すべきかが自明であるという利点もある．

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[9]，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手法を検討するために，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

⋆1 http://d.hatena.ne.jp/keyword/

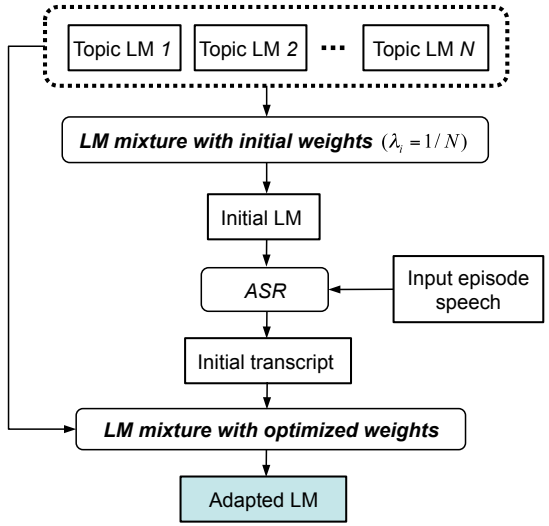

図 1 動的言語モデリング手法 (カジュアルな発話スタイルに対処するために，各トピック言語モデルは事前に話し言葉テキスト (CSJ) と線形補間を行っている．)

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

3.1 ポッドキャスト音声データ

本実験で利用するポッドキャスト音声データ[⋆1] の諸元について表 2 にまとめる．ここで，

⋆1 各ポッドキャストの番組名と PodCastle 上の URL は下記の通りである．実際の番組の URL も下記 URL から辿ることができる．ただし既に配信を終了した番組も存在する．
A:**「読売ニュース ポッドキャスト」**http://podcastle.jp/podcasts/show/14
B:**「聴くトク報知 」**http://podcastle.jp/podcasts/show/21
C:**「森本毅郎スタンバイ」**http://podcastle.jp/podcasts/show/9
D:**「伊藤洋一のビジネストレンド」**http://podcastle.jp/podcasts/show/42
E:**「５分でわかる証券基礎講座」**http://podcastle.jp/podcasts/show/293
F:**「吉田健康〜あなたのドクターたかよしです。」**http://podcastle.jp/podcasts/show/371
G:**「長谷川滋利の野球術」**http://podcastle.jp/podcasts/show/12
H:**「JUNK2 カンニング竹山 生はダメラジオ」**http://podcastle.jp/podcasts/show/107

表 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)

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．ただし学習セットの単語数は，実際の書き起こしではなく音声認識結果の単語数を示している．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．

3.2 音声認識システム

音声認識には，PodCastle 音声認識システムを用いた[8]．まず，ポッドキャストエピソードの音響ストリームを，GMM ベースの音響イベント検出器により音声，音楽，無音に分割する．ここで得られた各音声発話を下記のマルチパスデコーダにより認識を行う．

(1) まず，2-gram を用いた N-best 木構造ビームサーチにより単語グラフを生成する．次に，3-gram を用いて単語グラフをリスコアし，得られた単語仮説を用いて教師なし MLLR 適応[17] を行う．

(2) 適応された音響モデルを用いて，上記デコーディングを再度実行し単語グラフを再構築する．最後に，単語グラフに対して consensus デコーディング (単語誤り最小化デコーディング)[18] を実行し，confusion network を生成する．confusion network 中の最尤候補を抽出し，最終認識結果とする．

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．

表 3 動的言語モデリング手法 (全モデル混合) 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，“教師あり”，“教師なし” は混合重み最適化に正解の書き起こしを利用した場合，音声認識結果を利用した場合をそれぞれ示す.

ID	ベースライン	動的言語モデリング			
		w/o *podcast LM*		w/ *podcast LM*	
		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**

3.3 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 “initial LM”) を用いた際の認識性能である．ここでの動的言語モデリングは，2.3 節で述べた 2 つの混合手法のうち，**全モデル混合**を用いた場合の結果である．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，3.2 節で述べたように，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 “Initial trascript” で MLLR 適応した音響モデル) を用いた．

3.3.1 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 (w/o *podcast LM*) 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[11] 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．

表 4 全モデル混合と選択的モデル混合の比較

手法	単語誤り率 (%)	平均語彙サイズ	総未知語数
全モデル混合	33.7	286345	373
選択的モデル混合	34.2	87430	1169

3.3.2 全モデル混合と選択的モデル混合の比較

表 4 に，全モデル混合，選択的モデル混合それぞれを用いた場合の最終的な単語誤り率 (全エピソードの平均) を示す．混合重み自動推定は教師なしで行い，選択的モデル混合におけるトピック言語モデル選択の閾値は 0.1 とした．結果より，本実験では全モデル混合手法の方が，トピック言語モデルを絞り込む選択的モデル混合手法よりも高い性能を示した (絶対値で 0.5%)．この理由としては未知語の影響が主として挙げられ，選択的モデル混合では，語彙が特定トピックに絞り込める一方で，本評価セットにおいては未知語のカバー率が大きく減少していた．これは本研究で扱うポッドキャスト音声認識タスクが語彙，トピックともに大きな広がりを持ち，比較的大規模な言語モデルを適用する必要があることを示している．ただし，選択的モデル混合ではよりコンパクトなモデルを構成でき，低コストで音声認識を実行できる利点もあるため，今後，未知語を考慮した語彙選択手法を導入するなど更なる改善の余地がある．

3.3.3 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 “w/ *podcast LM*” 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動的言語モデリング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，本研究では，Web ニューステキストを有効活用することで，入力エピソードに対して動的にトピック適応を行う動的言語モデリング手法を構築した．提案手法では，25 のトピックにカテゴライズされた大規模なニュース記事データを用いてトピック言語モデルを学習し，入力エピソードごとに重みを最適化し，モデルレベル混合を行う．さらに，ポッドキャストの他のエピソードデータをもとに学習したポッドキャスト依存言語モデルを混合処理に組み込むことで，個々のポッドキャストエピソードのトピックにより大きく適応化していく．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 ニュースデータは，一般社会にお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[19]，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) Ogata, J. and Goto, M.: 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，Vol.25, No.1, pp.104–113 (2010).
4) Federico, M. and Bertoldi, N.: Broadcast news LM adaptation over time, *Computer Speech & Language*, Vol.18, pp.417–435 (2004).
5) Kawahara, T., Nanjo, H., Shinozaki, T. and Furui, S.: Benchmark test for speech recognition using the corpus of spontaneous japanese, *Proc. SSPR 2003* (2003).
6) Akita, Y., Mimura, M. and Kawahara, T.: Automatic Transcription System for Meetings of the Japanese National Congress, *Proc. of Interspeech 2009* (2009).
7) Lefevre, F., Gauvain, J.-L. and Lamel, L.F.: Genericity and portability for task-independent speech recognition, *Computer Speech & Language*, Vol.19, pp.345–363 (2005).
8) Ogata, J., Goto, M. and Eto, K.: Automatic Transcription for a Web 2.0 Service to Search Podcasts, *Proc. of Interspeech 2007*, pp.2617–2620 (2007).
9) 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).
10) Lei, X., Wu, W., Wang, W., Mandal, A. and Stolcke, A.: Development of the 2008 SRI mandarin speech-to-text system for broadcast news and conversation, *Proc. of Interspeech 2009* (2009).
11) Tur, G. and Stolcke, A.: Unsupervised language model adaptation for meeting recognition, *Proc. ICASSP2007* (2007).
12) Vergyri, D., Stolcke, A. and Tur, G.: Exploiting user feedback for language model adaptation in meeting recognition, *Proc. of ICASSP 2009* (2009).
13) Hsu, B.-J.P. and Glass, J.: Language model parameter estimation using user transcription, *Proc. of ICASSP 2009* (2009).
14) Meng, S., Thambiratnam, K., Lin, Y., Wang, L., Li, G. and Seide, F.: Vocabulary and language model adaptation using just one speech file, *Proc. ICASSP 2010* (2010).
15) 松原勇介，緒方　淳，後藤真孝：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 自然言語処理 2008-NL-185-6，pp.39–44 (2008).
16) Jelinek, F. and Mercer, R.L.: Interplated estimation of Markov source parameters from sparce data, *Proc. Workshop on Pattern Recognition in Practice* (1980).
17) Gales, M. J.F.: Maximal likelihood linear transformations for HMM-Based speech recognition, *Computer Speech & Language*, Vol.12, pp.75–98 (1998).
18) Mangu, L., Brill, E. and Stolcke, A.: Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Network, *Computer Speech & Language*, Vol.14, No.4, pp.373–400 (2000).
19) Hsu, B.-J. P.: Generalized linear interpolation of language models, *Proc. ASRU* (2007).