ベイズ推論を用いた連続音声からの言語モデル学習

Graham NEUBIG[†1]　三村正人[†1]
森信介[†1]　河原達也[†1]

本稿ではテキストを用いず，連続音声のみから言語モデルを学習する方法を提案する．音響モデルのみを用いて作成された音素ラティスに対して推論を行い，単語境界と言語モデルを同時に学習する．具体的には，ノンパラメトリックベイズ法に基づく階層的 Pitman-Yor 言語モデルを利用し，パラメータは WFST に基づいたギブスサンプリングで推定する．会議音声を用いた実験において，提案手法によって学習された言語モデルはパープレキシティ及び音声認識の音素誤り率を有意に改善することができた．さらに，ラティス処理と単語単位の獲得が誤り率の改善に貢献していることがわかった．

Learning a Language Model from Continuous Speech using Bayesian Inference

Graham Neubig,[†1] Masato Mimura ,[†1]
Shinsuke Mori [†1] and Tatsuya Kawahara [†1]

This paper proposes a technique for learning a language model directly from continuous speech, without the use of text. Inference is performed over phoneme lattices generated using only acoustic model scores, and word boundaries and a language model are learned simultaneously. A Bayesian non-parametric Hierarchical Pitman-Yor language model is used, and parameters are estimated with WFST-based Gibbs sampling. An experiment was performed using meeting speech, and language models built using the proposed techniques were able to significantly lower ASR phoneme error rates. In addition, lattice processing and word boundary discovery were shown to contribute significantly to this improvement.

1. はじめに

音声認識システムにおいて，言語モデルは音響的処理において生じる曖昧性を解消する役割を果たし，高い認識精度を実現するには必要不可欠である．通常，言語モデルの学習にはデジタル化されたテキストを用いて，認識対象のドメインや発話スタイルに合致した大規模の学習データが必要である．

しかし多くの場合，このようなデジタル化されたテキストは存在しないか，入手困難である．例えば，デジタル化されたテキストが非常に少ない，または全く存在しない言語や方言は数多い．また，話し言葉にはフィラーや口語的表現など，通常の書き言葉では見られない現象が存在し，テキストから学習された言語モデルを音声認識に利用するためにこれらの特徴を反映させなければならない[1)]．テキストではなく，音声のみから言語モデルを学習することができれば，デジタル化された学習テキストの必要がなくなり，話し言葉や方言等を含んだ言語モデルを直接学習することが可能となる．このため，テキストを利用せずに連続音声から言語モデルを学習する可能性を探求することは，非常に興味深い研究課題である．

本稿では，連続音声データと音響モデルのみを用いて，教師なしで言語モデルと単語辞書を学習する手法を提案する．具体的には，ノンパラメトリックベイズ法に基づく階層的 Pitman-Yor 言語モデル（HPYLM）を利用した単語分割法[2)] を連続音声に対して適用する．システムの各部分を重み付き有限状態トランスデューサ（WFST）で表現し，WFST 合成により作成されたラティスに対するギブスサンプリングによってベイズ推論を行う．これにより，音響的曖昧性を反映させた音素ラティスに対して学習が可能となる．

2. 教師なし単語分割

教師なし単語分割は人間の言語獲得過程の解明や，音声認識のための最適な単位の発見を目的として研究されてきた[2)–6)]．特に，近年では統計的モデルを利用する手法が主流となっている．ほとんどの場合では日本語や中国語などの分かち書きされない言語や，単語境界のない音素書き起こしなどのテキストに対して学習を行い，テキストの分割精度やエントロピーを基準として評価されている．

通常これらの手法は，コーパス $\mathcal{X}$ に含まれている文字列 $\mathbf{x} = x_1, \ldots, x_I$ がある言語モデル G によって生成されたと仮定する．可能な言語モデル G に対して事前分布 $P(G)$ を定義し，最大事後確率（MAP）推定やベイズ推論を用いて，モデルとデータの同時確率が高く

†1 京都大学 情報学研究科
Kyoto University, Graduate School of Informatics

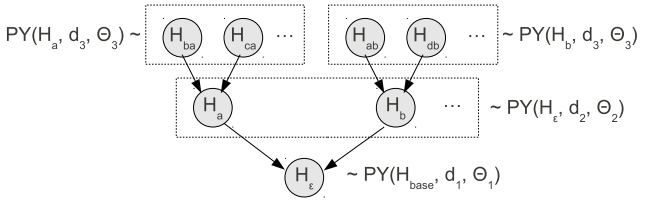

図 1 3-gram の HPYLM の 1 例．各ノードは 1 つの n-gram 分布を表し，そのノードの名前は分布の履歴を表す．枠に囲まれているノードは同一の基底測度を共有する．

なる解を探索する．

$$P(\mathcal{X}, G) = P(\mathcal{X}|G)P(G) \tag{1}$$

2.1 階層的 Pitman-Yor 言語モデル（HPYLM）

本稿では Mochihashi ら[2] の階層的 Pitman-Yor 言語モデル（HPYLM）による教師なし単語分割法を拡張し，連続音声から言語モデルを学習する．HPYLM は Pitman-Yor 過程という確率過程に基づいた言語モデルである[7]．Pitman-Yor 過程は基底測度 H，ディスカウント d，強さ θ という 3 つのパラメータを持つ．H は生成されうる分布の期待値を表し，d と θ は未観測の事象の出現確率を調整するスムージング係数を表す．

HPYLM は図 1 に示す通り，Pitman-Yor 過程を階層的な形にしたものである．n 番目の階層のノードは履歴長 $n-1$ の n-gram の分布を表し，$n-1$ 番目の階層の分布を基底測度とする．図 1 からわかるように，複数の分布（例えば H_{ba} と H_{ca}）は同一の基底測度（H_{a}）を共有している．これは通常の平滑化された言語モデルで $P(w_i|w_{i-2}=b, w_{i-1}=a)$ と $P(w_i|w_{i-2}=c, w_{i-1}=a)$ はいずれも $P(w_i|w_{i-1}=a)$ と補間されることに相当する．また，各階層で共通のディスカウント d_i と強さ θ_i を共有する．

まとめると，HPYLM は以下の式で表される．

$$LM \sim HPY(H_{base}, \mathbf{d}_1^n, \theta_1^n). \tag{2}$$

H_{base} は H_ϵ の基底測度であり，通常では語彙内の各単語に対する一様分布とする．

2.2 HPYLM による教師なし単語分割

教師なし単語分割において HPYLM を利用することにより，式 (1) に現れる言語モデルの事前確率 $P(G)$ を式 (2) の HPYLM 生成確率と設定することができ，モデルの複雑さと表現力のバランスを取ることができる．

まず，HPYLM で単語分割を行うために，コーパス中の文字列が独立同分布であり，階層的 Pitman-Yor 言語モデル LM によって生成されたと仮定する．言語モデル LM は表層の文字列 $\mathbf{x}$ を直接生成するのではなく，未観測の単語列 $\mathbf{w}$ を生成したとする．ただし，単語列 $\mathbf{w}$ の文字をつなげた時（関数 $ct(\mathbf{w})$ で表す）文字列 $\mathbf{x}$ が復元できる．したがって，$P(\mathcal{X}|LM)$ は以下の分布となる：

$$\begin{aligned} P(\mathcal{X}|LM) &= \prod_{\mathbf{x}\in\mathcal{X}} P(\mathbf{x}|LM) \\ &= \prod_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{w}\in\{\tilde{\mathbf{w}}:ct(\tilde{\mathbf{w}})=\mathbf{x}\}} P(\mathbf{w}|LM). \end{aligned}$$

LM が式 (2) のような階層的 Pitman-Yor 過程によって生成されたとすると，LM は未観測の単語列を生成し，各単語とその単語を構成する文字の関係を表すモデルが必要となる．このため，LM の基底測度 H_{base} を一様分布ではなく，文字 n-gram の未知語モデル SM の単語生成確率とする．これにより，各単語の表記に確率を与え，単語と表記を確率的に結びつけることができる．未知語モデル SM の基底測度として，コーパス中の各文字に対する一様分布を利用する．コーパス全体の生成過程は以下の通りである：

$$\begin{aligned} SM &\sim HPY(U, \mathbf{d}_{SM}, \theta_{SM}) \\ LM &\sim HPY(SM, \mathbf{d}_{LM}, \theta_{LM}) \\ \mathcal{X} &\sim LM. \end{aligned}$$

簡潔にするため，これ以降言語モデル LM と未知語モデル SM を併せて G と記述する．

この生成モデルを利用して任意の文字列 $\mathbf{x}$ に出現確率を与えるためには，以下のような，G を周辺化した予測分布を計算する必要がある．

$$P(\mathbf{x}|\mathcal{X}) = \int_G \sum_{\mathbf{w}\in\{\tilde{\mathbf{w}}:ct(\tilde{\mathbf{w}})=\mathbf{x}\}} P(\mathbf{w}|G)P(G|\mathcal{X})dG.$$

しかし，上記の関数を直接計算することは困難であり，計算を実現するためにいくつかの独立仮定を置く．まず，Viterbi 近似に基づき，$\mathbf{x}$ の出現確率は $ct(\mathbf{w})=\mathbf{x}$ の条件を満たす単語列 $\mathbf{w}$ の中で，最も尤度の高い $\mathbf{w}$ の確率と同等であると仮定する．また，ギブスサンプリングを用いて $P(G|\mathcal{X})$ から S 個のサンプルを生成し，$P(\mathbf{w}|G_s)$ の平均で実際の事後確率分布を近似する．

$$P(\mathbf{x}|\mathcal{X}) \approx \frac{1}{S}\sum_{s=1}^{S} \max_{\mathbf{w}\in\{\tilde{\mathbf{w}}:ct(\tilde{\mathbf{w}})=\mathbf{x}\}} P(\mathbf{w}|G_s). \tag{3}$$

Mochihashi らは，サンプリングの効率化を図るために forward filtering-backward sampling という手順を用いて文ごとにサンプリングを行っている．

2.3 forward filtering-backward sampling

forward filtering-backward sampling は隠れマルコフモデル (HMM) の前向き後ろ向き

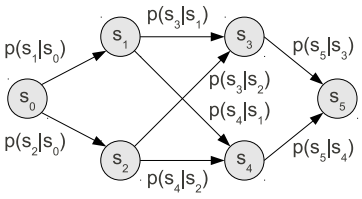

図 2 WFSA の一例

アルゴリズムに類似したものであり，重み付き有限有限オートマトン (WFSA) のパスを効率的にサンプリングすることを可能とする[8)⋆1]．本節は図 2 を例としてこの手順を説明する．

まず，forward filtering の段階で，確率は初期状態から逐次的に計算する．初期状態 s_0 の前向き確率を $f_0 = 1$ と定義すると，以下の状態の前い向き確率は以下のように計算される．

$$f_1 = p(s_1|s_0) * f_0$$
$$f_2 = p(s_2|s_0) * f_0$$
$$f_3 = p(s_3|s_1) * f_1 + p(s_3|s_2) * f_2$$
$$\vdots$$

backward sampling の段階で，前向き確率と各辺の遷移確率に基づいて，受理状態から初期状態へと通るパスをサンプリングする．例えば，s_5 につながる辺は以下のような確率分布に基づいてサンプリングする：

$$P(s_4 \to s_5) \propto P(s_5|s_4) * f_4$$
$$P(s_3 \to s_5) \propto P(s_5|s_3) * f_3$$

この過程を経て，WFSA 中のすべてのパスの中から，1 個のパスを確率的にサンプリングすることができる．言い換えると，ある確率分布を WFSA として表現できれば，その分布からサンプリングできることになる．

2.4 WFST による教師なし単語分割のサンプリング

Mochihashi ら[2)] は forward filtering-backward sampling で HPYLM に基づいた単語分割のパラメータを推定している．モデルとマルコフ過程の関係について言及してはいるが，実装の段階では明確に WFSA として定式化せずに，手続きを述べているのみである．ここでは，Mochihashi らのアルゴリズムを，重み付き有限状態トランスデューサ (WFST[9)]) の合成とサンプリングとして定式化する．これにより，WFST の合成に標準的なライブラリ[10)] を利用できるため実装が比較的容易となり，次節で述べる連続音声の単語分割も可能となる．

⋆1 ただし，グラフに循環が含まれておらず，辺の重みが遷移確率を表している場合に限る．

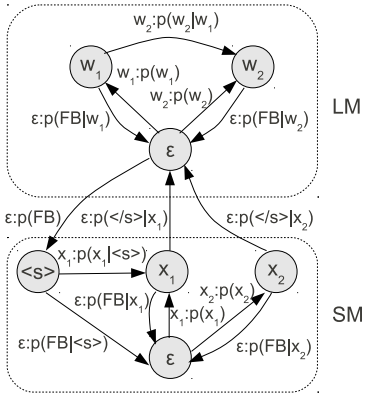

図 3 G を表現する WFST．“FB” はより短い履歴へのバックオフを表し，“<s>” と “</s>” はそれぞれ SM の初期状態と終了記号を表す．

サンプリングする WFSA を構築するために，以下の 3 つの WFST を構築する：

- X：入力文字列 $\mathbf{x}$ を表す WFSA
- L：$\mathbf{x}$ を入力とし，全ての可能な分割候補 $\mathbf{w}$ を出力とする単語辞書 WFST
- G：単語列 $\mathbf{w}$ に言語モデル確率 $P(\mathbf{w}|G)$ を付与する言語モデル WFSA

X の構築法は自明であり，L と G の構築法は文献[9)] で詳しく述べられている．しかし，通常の音声認識は未知語のモデル化を行わないため，言語モデルに SM を利用せずに，LM のみで表現する．このため，LM と SM の両方を利用する G は通常の言語モデル WFST とは多少異なる．

LM と SM を組み合わせた WFST を図 3 に示す．ここでは，LM の履歴なしの状態から SM の初期状態へとバックオフする辺と，SM の終了記号</s>を受理し，SM から LM へ戻る辺を導入することで，1 つの WFST で G を表現する．

$X \circ L \circ G$ の順に 3 つの WFST を合成して得られた WFSA のパスは，言語モデル G の重みがついた分割候補を表す．この WFSA に対して forward filtering-backward sampling を行うことで，分割された文字列を得ることができる．これを文献[2)] の文サンプリング手順と置き換えることによって，学習枠組み自体を変更することなく同等の結果が得られる．

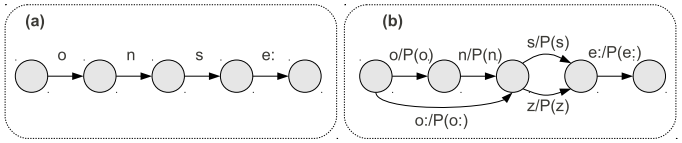

図 4 **(a)** テキストの X，**(b)** 連続音声の X

3. 連続音声からの言語モデル学習

3.1 関 連 研 究

テキストに対する教師なし単語分割の研究が数多くなされているにもかかわらず，音声を用いた単語分割や語彙獲得の研究は比較的少ない．その少ない例として，1-best 音素認識結果から単語境界や言語モデルを学習する方法や[3),11)]，マルチモーダルデータの相互情報量で単語境界の曖昧性を解消する方法[12)]，音響特徴類似度によるマッチングで音声の類似している箇所を発見する方法[13)] などがある．

本稿の提案手法は，2 つの点で先行研究と大きく異なる．まず，多くの先行研究が 1-best の認識結果を利用しているのに対して，提案手法では音素ラティスを用いて学習を行う．これにより，音響的処理の曖昧性による認識誤りを吸収することができ，より正確なモデル学習が期待できる．また，本手法では単語境界のみならず，n-gram モデルで表現される統語情報も学習することができる．この言語モデルは，学習データ以外の発話の音声認識にも利用可能である．さらに，Goldwater ら[5)] の実験により，n-gram モデルなどで得られる文脈情報は，高い分割精度を実現するには必要不可欠であることが示されている．

3.2 連続音声のための HPYLM サンプリング

我々が 2.4 節で提案した WFST による定式化により，HPYLM を用いた単語分割を容易に連続音声に適用することができる．実際には，テキストの単語分割で一意の文字列を表す X を，WFST に基づく音声認識で音響モデルを表す HMM と置き換えれば，2.4 節で述べた手続きを用いて連続音声に対して単語分割と言語モデル学習を行うことが可能となる．

しかし，HMM をそのまま利用すると，音響モデルが許す全ての音素列が forward filtering-backward sampling で展開する仮説空間となり，全探索は困難となる．ここでは，全音素列の展開を避けるために，学習の前に予め音響モデル確率の低い仮説を枝刈りし，図 4 で示すように音響モデル確率の高い仮説の音素ラティスを X として利用する．

4. 実 験 評 価

国会審議の音声を用いた実験で提案手法の有効性を検証した．国会審議は多くの先行研究が使っている対幼児音声[12)] と比較して，語彙が大きく，平均的な発話単位が長い．

4.1 実 験 設 定

まず，トライフォンの音響モデルで音声から音素ラティスを作成した[⋆1]．音素ラティス作成時に，385 音節に一様分布を与える言語モデルを利用した．この 385 音節は日本語話し言葉に用いられる音節を網羅し，「リェッ」や「ビャー」など日本語にほとんど出現しない音節も多く含まれている[⋆2]．

学習データの大きさを 119〜1904 発話（7.9〜116.7 分）の間で変動させ，学習データに含まれない 500 発話（27.2 分）をテストセットとした．学習された言語モデルを用いて，テストセットの音素ラティスのリスコアリングを行い，リスコアリング後の音素誤り率 (PER) を評価基準とする．ある程度主観に基づく単語分割精度と異なり，PER は獲得された言語モデルの客観的な評価基準である．

言語モデルの学習を行わずに，音節 0-gram モデルで認識を行った場合の PER は 34.20%であり，音素ラティスの最適なパスを選んだ場合の PER（オラクル誤り率）は 8.10%であった．この結果から，完全な言語モデルを獲得できたとしても多くの誤りが残ることがわかる．

学習コーパスに対して計 70 回のサンプリングを行ない，最初の 20 回を burn-in として捨て，次の 50 回から G のサンプリングを行った．より早く真の分布に収束させるために，最初の 10 回に対して文献[5)] のアニーリング法を適用した．また，言語モデルの対数線形重みは 5 に設定した[⋆3]．式 (3) では $P(\mathbf{x}|G)$ を $ct(\mathbf{w}) = \mathbf{x}$ となる最も高い単語列確率で近似するが，50 個の分割単位の異なる言語モデルによる確率を組み合わせてこの解を探索する方法は自明ではない．そのため，この組み合わせの近似として，個別にすべてのモデルを用いて最も確率の高い解を求め，それぞれの解を ROVER で組み合わせた[16)]．

4.2 n-gram による文脈情報の効果

まず，n-gram のオーダーを変動させて，学習過程に文脈情報を利用する効果を調べた．SM のオーダーを 3 と設定し，LM のオーダーとして 1-gram，2-gram，3-gram をそれぞれ利用した．それぞれのモデルの音素誤り率を図 5 に示す．

⋆1 事前に学習された音響モデルに依存するため，提案手法は完全に教師なしであるとは言えないが，教師なし・言語非依存の音響モデル学習を扱う先行研究[14),15)] はあり，これらと本手法の組み合わせはこれからの課題である．
⋆2 音節単位のデコーディングは音声認識デコーダの制限によるもので，本質的に本手法で必要であるわけではない．
⋆3 予備実験では 5〜10 の間の値はほぼ同等の精度となった．

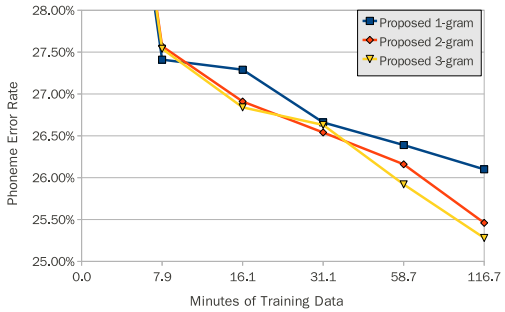

図 5 n-gram 長と音素誤り率の関係

表 1 116.7 分の学習データを用いたモデルの語彙サイズと n-gram 数

	1-gram	2-gram	3-gram
語彙サイズ	4480	1351	708
LM のエントリー数	4480	16150	38759
SM のエントリー数	9624	3869	2426

まず，言語的情報を利用しない 0-gram モデルと比較して，7.9 分の音声のみから学習されたモデルでも 7%を超える PER の改善が見られ，116.8 分では 8.92%の絶対改善が見られた．また，データサイズが大きくなるほど 2-gram と 3-gram の精度が 1-gram を上回り，連続音声からでも文脈を表現するモデルが学習可能であることがわかった．

Goldwater ら[5] が観測したように，1-gram モデルは比較長い単位を学習し，単語ではなく句に相当する単位を獲得する．言語モデルの学習が終わった後の語彙サイズ，n-gram 数を表 1 に示す．1-gram モデルの語彙は 2-gram と 3-gram モデルの語彙より大きく，統語情報に相当する n-gram 数はより小さい．つまり，LM で統語情報が利用できない代わりに，SM で語彙を拡大し，モデルの表現力を増やしている．

4.3 他の言語モデル構築法との比較

本稿の提案手法である音素ラティス処理を，異なる 3 つの言語モデル構築法と比較した．まず，提案手法を音素ラティスではなく，音節認識の 1-best 結果で学習した．これにより，1-best 処理を利用した先行研究との比較ができ，ラティス処理の有効性を調べることができる．さらに，単語境界の学習の必要性を検証するために，単語境界の学習を行わない，音節 3-gram モデルを音節認識の 1-best 結果で構築した．

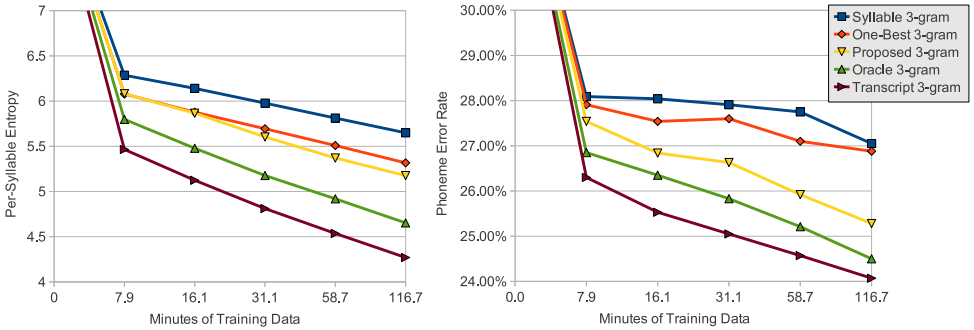

図 6 提案手法のラティス処理，1-best 処理を用いたモデル，音節ごとの言語モデル，音素ラティスのオラクルパスを用いたモデル，忠実な書き起こしを用いたモデルの性能比較

最後に，テキストを必要とする通常の言語モデル学習との比較を行うために，人間による正確な書き起こしを用いて言語モデルを構築し，比較を行った．単語分割と発音推定を自動的に付与し，未知語の発音は人手で付与した．この分割された単語の音素列に対して 3-gram の LM と SM を学習し，補間 Kneser-Ney 法で平滑化を行った．

4 通りの学習法の PER とエントロピーを図 6 に示す．PER については，提案手法のラティス処理で構築したモデルは，1-best の認識結果で構築したモデルを大きく上回り，複数の音素認識仮説を利用することがモデルの性能向上につながることを確認することができた．また，1-best 結果の中で，音節 3-gram モデルより，単語境界を獲得したモデルの方が低い PER となることから，語彙獲得は言語モデルの性能向上につながることも確認できた．

エントロピーの評価では，提案手法が 1-best を用いたモデルをわずかに上回った．エントロピーの差が PER の差より小さい理由として，体系的な発音の怠けが考えられる．例えば，正解の書き起こしで「カンガエテオリマス」と書かれている箇所の「テオ」は発音の怠けにより「ト」となることが多い．これにより学習された語彙に「カンガエトリマス」という項目が入っている．正解データと比べて 1 つの音素のみが異なるため，PER に大きな影響を与えることはないが，厳密な単語のマッチが必要なエントロピー評価を悪化させる原因となる．しかし，学習された単語は正しく発音の怠けを表しているのであれば大きな問題ではなく，むしろ望ましいことであるとも言える．

人手による書き起こしには音素ラティスに存在する曖昧性や誤りが含まれていないため，書き起こしで学習したモデルは提案手法のモデルより高い精度を実現した．しかし，提案手法のモデルは 1-best の認識結果で構築したモデルと異なり，データが増えるとともに人手

による書き起こしで構築したモデルとほぼ同等の改善率が見られた．

人手による書き起こしのモデルの精度に到達できなかった理由の 1 つとして，音素ラティスの比較的高いオラクル誤り率（8.10%）が考えられる．この影響を検証するために各音素ラティスの中から最も PER の低いパスを選び，これを言語モデルの学習に利用した．図 6 の通り，オラクルを用いたモデルの PER は提案手法と書き起こしを用いたモデルの中間となり，約半数の誤りはオラクル誤り率に起因するものであると考えられる．

5. おわりに

本稿では，テキストを用いずに連続音声のみから言語モデルを学習する手法を提案し，実験で提案手法による言語モデル学習が可能であることを示した．特に，単語単位と言語モデルの同時学習や，ラティス処理による複数仮説の考慮が言語モデルの精度向上につながることがわかった．

提案手法は様々な分野で新たな研究課題と可能性を開くものと考えられる．テキストコーパスが全く，または十分に存在しない言語や方言に対して本手法を適用することで，音声のみから語彙と言語モデルを学習することができる．また，提案手法を半教師ありの枠組みで利用することで，テキストコーパスから学習した言語モデルを新しい分野・話者・方言に適応することが可能になる．

今後の技術課題として，より大規模なデータへの適用がある．現在では，音素ラティスを用いたサンプリングはリアルタイムの数十倍のオーダーの時間を要し，単独の計算機ではさらに大規模なデータへの展開が困難である．しかし，近年ではサンプリングの並列化技術が進んでおり，複数のマシンを利用することでさらに大きなデータが取り扱い可能となる[17]．また，音素ラティスのオラクル誤り率はモデルの精度に大きく影響していることから，X を音素ラティスではなく，音響モデルの HMM で直接計算することによってさらなる精度向上が期待できる．ビームサンプリング[18] などの技術を利用し，仮説空間を枝刈りしながらサンプリングを行うことで HMM との統合は可能となると考えられる．

謝辞　本稿の内容についてご意見頂いた持橋大地氏に感謝の意を表する．

参考文献

1) Akita, Y. and Kawahara, T.: Topic-independent Speaking-style Transformation of Language Model for Spontaneous Speech Recognition, *Proc. ICASSP2007*, pp. 33–36 (2007).
2) Mochihashi, D., Yamada, T. and Ueda, N.: Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Modeling, *Proc. ACL09* (2009).
3) de Marcken, C.: The Unsupervised Acquisition of a Lexicon from Continuous Speech, Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA (1995).
4) Brent, M.R.: An efficient, probabilistically sound algorithm for segmentation and word discovery, *Machine Learning*, Vol.34, No.1, pp.71–105 (1999).
5) Goldwater, S., Griffiths, T.L. and Johnson, M.: A Bayesian framework for word segmentation: Exploring the effects of context, *Cognition*, Vol.112, No.1, pp.21–54 (2009).
6) Hirsimaki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S. and Pylkkonen, J.: Unlimited vocabulary speech recognition with morph language models applied to Finnish, *Computer Speech & Language*, Vol.20, No.4, pp.515–541 (2006).
7) Teh, Y.W.: A hierarchical Bayesian language model based on Pitman-Yor processes, *Proc. ACL06*, pp.985–992 (2006).
8) Scott, S.: Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century., *Journal of the American Statistical Association*, Vol.97, No.457, pp.337–352 (2002).
9) Mohri, M.: Finite-state transducers in language and speech processing, *Computational Linguistics*, Vol.23, No.2, pp.269–311 (1997).
10) Allauzen, C., Riley, M., Schalkwyk, J., Skut, W. and Mohri, M.: OpenFst: a general and efficient weighted finite-state transducer library, *Proc. CIAA '07*, pp.11–23 (2007).
11) Gorin, A., Petrovska-Delacretaz, D., Riccardi, G. and Wright, J.: Learning spoken language without transcriptions, *Proc. ASRU99* (1999).
12) Roy, D. and Pentland, A.: Learning words from sights and sounds: A computational model, *Cognitive Science*, Vol.26, No.1, pp.113–146 (2002).
13) Park, A. and Glass, J.: Unsupervised pattern discovery in speech, *IEEE Transactions on Audio Speech and Language Processing*, Vol.16, No.1, p.186 (2008).
14) Schultz, T. and Waibel, A.: Language-independent and language-adaptive acoustic modeling for speech recognition, *Speech Communication*, Vol.35, No.1, pp.31–52 (2001).
15) Lamel, L., Gauvain, J. and Adda, G.: Lightly supervised and unsupervised acoustic model training., *Computer Speech and Language* (2002).
16) Fiscus, J.: A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER), *Proc. ASRU97* (1997).
17) Asuncion, A., Smyth, P. and Welling, M.: Asynchronous distributed learning of topic models, *Proc. NIPS08*, Vol.21 (2008).
18) Van Gael, J., Saatci, Y., Teh, Y. and Ghahramani, Z.: Beam sampling for the infinite hidden Markov model, *Proc. ICML08*, ACM, pp.1088–1095 (2008).