Vol.2009-SLP-77 No.8
2009/7/17

統計的言語モデル変換を用いた音響モデルの準教師つき学習

三村正人[†1] 秋田祐哉[†1] 河原達也[†1]

話し言葉音声認識における学習データ量不足の問題に対処するために、人手による忠実な書き起こしが存在しない条件で、音響モデルの準教師つき学習を行うアプローチが検討されている。本報告では、この準教師つき学習における効果的なラベルの作成手法について提案する。提案手法では、会議録のテキストデータに統計的話し言葉変換を適用して会議の詳細な単位 (ターン) 毎に制約の強い言語モデルを作成し、この言語モデルを用いて音声認識を行うことで音響モデル学習のためのラベルを作成する。評価実験により、従来手法よりも高い精度のラベルを作成できること、またこのラベルを用いて人手のラベルを用いた場合と同等の精度のモデルが学習できることを示す。

Lightly Supervised Training of Acoustic Model Using Language Model Transformatiion

Masato Mimura,[†1] Yuya Akita[†1]
and Tatsuya Kawahara[†1]

To address the problem of insufficient amount of training data for spontaneous speech recognition, the approach of lightly supervised training of the acoustic model has been investigated. In this report, we propose an efficient method to provide a constrained and compact language model for generating transcripts using the language model transformation scheme. The experimental evaluation demonstrated that the proposed scheme can generate acculate labels and realizes the comparable ASR performance to the case using manual transcripts.

†1 京都大学 学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University

1. はじめに

近年、大語彙連続音声認識の主要な対象は、講演や会議といった話し言葉音声に移行しつつある。話し言葉音声は読み上げ音声では見られないような非流暢な現象を伴うため、人手による忠実な書き起こしの作成には高いコストがかかり、大規模なコーパスを構築することが困難である。したがって、音響モデルの学習を行う際のデータ量の不足が問題となる。この問題に対処するために、音響モデルの準教師つき学習 (lightly supervised training) のアプローチが研究されている[1)]。このアプローチでは、発話の忠実な書き起こしの代わりに、低コストで利用できるようなテキストデータを知識源として音響モデルの学習を行う。

Lamel ら[1)]は、ニュース音声を対象として、字幕 (closed caption) のテキストデータを用いた準教師つき学習を提案している。多くの放送には字幕が付与されているが、字幕自体は実際の発話と異なるため、音響モデル学習のためのラベル (音素列またはこれを生成できる書き起こし) として直接利用することはできない。そこで、字幕のテキストデータから学習した言語モデルを用いて音声認識を行うことでラベルを作成する。また、ニュース音声には音楽や CM などの非音声区間が多数存在するため、認識結果と字幕を再度照合し、合致する区間のみをフィルタリングすることが効果的と報告している。Chan ら[2)]は、字幕に現れない表現にも対応するために、字幕モデルと別途構築したベースライン言語モデルを、前者に大きな重みをかけて合成し (biased language model)、音声認識を行っている。また、作成されたラベルを用いた学習データの追加により、通常の ML 学習だけでなく、識別学習 (MPE[3)]) においても認識精度が向上することを報告している。

Paulik ら[4)]は、欧州議会音声を対象として、会議録のテキストを知識源として利用している。欧州議会の本会議では、発言はほぼ読み上げ調であり、会議録と実際の発話との相違は小さいと考えられる。報告では、会議録のテキストから直接構築した言語モデルにより音声認識を行い、有効なラベルを作成している。また、当該会議のテキストに大きな重みをかけて学習した制約の強いモデルを用いることで、高い精度のラベルが得られている。

我々は衆議院審議音声の自動書き起こしシステムの研究開発を行っている。日本の国会においても欧州議会と同様に会議録が利用可能であるが、委員会が会議の大半を占めるため、よりインタラクティブで自発的な発話が主となり、実際の発話内容と会議録の相違が大きい。国会審議音声における発話の忠実な書き起こしと会議録の例を図 1 に示す。会議録では助詞「が」の挿入や、「いー」等のフィラーの除去による整形が行われている。

本報告では、この会議録をもとに、音響モデルの準教師つき学習のための効果的なラベル

Vol.2009-SLP-77 No.8
2009/7/17

発話

総理 おっしゃっ た とおり 、 これ は 、 我が国 いー 、 に のみ ならず 、 韓国 、 周辺 国 、 うー アジア 、 あー この 地域 全体 にとって 大きな 脅威 で あり ます

会議録

総理がおっしゃったとおり、これは、我が国のみならず、韓国、周辺国、アジア、この地域全体にとって大きな脅威であります 。

図 1 発話と会議録のテキストの例

の作成手法を提案し、人手の書き起こしが存在しない場合でも音響モデルの学習・更新が可能であることを示す。会議録をもとにラベルを作成する際は、上記の話し言葉特有の現象に対応することが課題となる。我々は、話し言葉 (発言の忠実な書き起こし) と整形済み文書 (会議録) との対応づけコーパスから、言語モデルのスタイル変換のための統計モデルの枠組みを提案し、音声認識用の言語モデルにおいて有効性を示している[5)]。本報告の提案手法では、この統計的言語モデル変換を個々の会議録に適用して話し言葉スタイルの言語モデルを構築し、この制約の強いモデルを用いて当該会議の音声認識を行うことによりラベルを作成する。

以下、2 章で言語モデルの統計的スタイル変換について概説し、3 章ではそれを用いて音響モデルの準教師つき学習を行う手法について述べる。4 章では提案手法を衆議院審議音声で評価した結果について述べる。5 章でまとめを行う。

2. 言語モデルの統計的話し言葉変換

言語モデルの統計的スタイル変換では、統計的機械翻訳の枠組みに基づき、話し言葉スタイル V と文書スタイル W の変換を行う。原理的には、この変換は双方向的である。すなわち、話し言葉の忠実な書き起こしを文書スタイルへ整形する方向へも、文書スタイルのテキストから書き起こしを復元する方向へも適用できる。デコードは統計的機械翻訳の枠組みに従い、以下のベイズ則に基づいて行われる。

$$p(W|V) = \frac{p(W) \cdot p(V|W)}{p(V)} \tag{1}$$

$$p(V|W) = \frac{p(V) \cdot p(W|V)}{p(W)} \tag{2}$$

なお、各式の分母は通常無視される。

ここで、(2) 式により V を一意に決定することは、(1) 式により整形を行う過程よりもはるかに難しい点に留意する。例えば、(2) 式においてフィラーはランダムに挿入されうるため、変換には任意性が存在するが、(1) 式においてはフィラーは確率 1 で除去される。したがって、V を一意に復元することよりも、次式のように V 上の統計的言語モデルを推定することの方が有意義であると考えられる。

$$p(V) = p(W) \cdot \frac{p(V|W)}{p(W|V)} \tag{3}$$

ここで重要なことは、文書スタイルのテキスト W は話し言葉の書き起こし V よりも豊富に存在する点である。すなわち、(3) 式に従えば、豊富な文書スタイルのテキスト W を用いて話し言葉音声認識のための言語モデル $p(V)$ を頑健に推定できる。

実際の変換は、次式のように N-gram カウントを操作することで行なわれる。

$$N_{gram}(v_1^n) = N_{gram}(w_1^n) \cdot \frac{p(v|w)}{p(w|v)} \tag{4}$$

v および w は、各スタイルにおける変換パターンである。これにより、置換 $w \to v$、w の脱落、v の挿入を、文脈を考慮してモデル化する[⋆1]。

条件つき確率 $p(v|w)$ および $p(w|v)$ は、忠実な書き起こしと文書スタイルテキストの対応づけコーパスから推定される。我々はこの対応づけコーパスを国会会議録の一部を用いて構築済みである。条件つき確率はコーパス中の各パターンの出現回数から推定される。

より適切なモデルとなるように変換パターンの隣接単語も考慮する。例えば、フィラー「あー」は、$\{w = (w_{-1}, w{+}1) \to v = (w_{-1}, \text{あー}, w_{+1})\}$ のようにモデル化される。また、品詞情報を用いたスムージングも行う。詳しい実装と評価については、文献[5)]を参照されたい。

我々はこの統計的言語モデル変換の枠組みを国会審議音声認識におけるベースライン言語モデルの構築に適用し、有効性を示している。

⋆1 通常の統計的機械翻訳と異なり、語順の入れ換えについては考慮しない

Vol.2009-SLP-77 No.8
2009/7/17

3. 音響モデルの準教師つき学習

本研究では、前章で述べた言語モデルの統計的話し言葉変換を、音響モデルの準教師つき学習に適用する。

国会では、すべての審議音声が収録されている。これらの音声に対して、忠実ではないがすべての発言内容を書き起こした会議録[⋆1]が作成されている。したがって、会議録をもとにラベルを自動で作成できれば、すべての音声データが直ちに音響モデルの学習データに利用できることになる。

前章で述べたように、会議録から書き起こしを一意に復元することは不可能であるが、会議録のテキストから (3) 式に従って話し言葉スタイルの統計的言語モデルを推定することができるため、このモデルを用いた音声認識を行うことにより、高い精度のラベルが作成できると期待できる。

提案手法の処理の流れを図 2 に示す。

3.1 言語モデルの作成

会議録は予算委員会、法務委員会等の委員会と本会議の各号の会議毎に作成される。各発言には発言者 ID が付与されており、それに従ってターン毎のテキストに分割できる。会議はおよそ 2 時間から 5 時間の長さであり、ターンは単一話者からなる 10 秒から 3 分程度 (平均 1 分 20 秒) の長さの区間である。

言語モデルがより強い制約となるためには、多くの話者や話題からなる会議全体でなく、ターン毎にモデルを作成することが重要である。提案手法は、ベースライン言語モデルとの補間を行う手法と異なり、個々のモデルのサイズが大きくならないので、ターンのような詳細な単位毎に言語モデルを用意することが可能である。

各ターンの会議録のテキストデータに、形態素解析を適用し、単語列および各単語の読みの表層形を取得する。次に、単語列を用いて N-gram エントリの抽出と出現回数のカウントを行い、(4) 式に従って話し言葉スタイルに変換する。変換された N-gram カウントを用いて、ターンに依存した話し言葉スタイルの言語モデルを推定する。ここで、会議録の情報を最大限利用するために、N-gram エントリのカットオフは行わない。

従来手法との比較について述べる。字幕や会議録の整形文のみを用いる手法[1)4)]では、音声に出現する話し言葉固有の表現に対処できず、また整形文を用いたフィルタリングにより

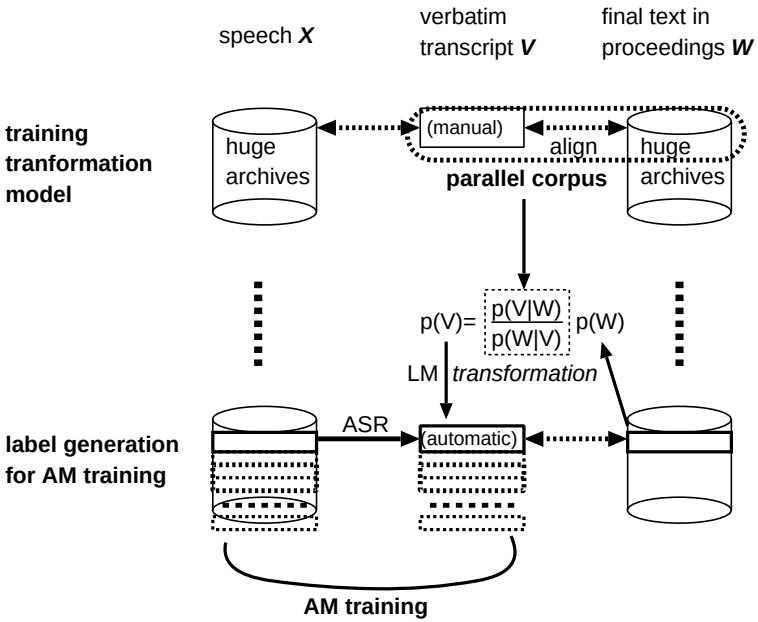

図 2 提案手法の概要

自発的な発話の区間はすべて除去されてしまう。また、ベースライン言語モデルと整形文から作成した言語モデルを重みつき混合して biased LM を作成する手法[2)]では、話し言葉スタイルの表現に対しては頑健な推定が可能だが、種々のドメインが混合しているため、当該音声に対して強い制約とはならない。提案手法は、話し言葉向け biased LM を作成するという点では従来手法と同一のアプローチだが、整形文のみから直接話し言葉スタイルの言語モデルを推定するため、コンパクトな大きさになる点に加えて、他ドメインの内容語が混入せず、より強い制約となる。

なお、音響モデルの学習を行うためには、最終的に音素列が必要となる。会議録のテキストから形態素解析を介して得られる読みは表層形のみである。特に話し言葉では発音変動が生じやすく、一つの表層形が複数の発音形を取りうるため、統計的な枠組みによりそれらを確率つきで予測し[6)]、認識辞書に追加しておく。

⋆1 http://kokkai.ndl.go.jp/

3.2 音声認識を用いたラベルの作成

提案手法により作成した言語モデルと認識辞書を用いて、対応するターンの音声の認識を行う。認識は、Julius[⋆1]のショートポーズセグメンテーションに基づく逐次デコーディングにより行う。逐次デコーディングの出力は、単語レベルおよび音素レベルの認識結果が付与されているとともに、ポーズにより最長 30 秒程度の扱いやすい区間に分割済みである。以降の学習はこの区間 (発話) を単位として行う。なお、この際の音響モデルは人手による忠実な書き起こしが存在する過去の衆議院審議音声データで学習したベースラインモデルを用いる。

また、認識結果を再度会議録と比較することで、誤りの訂正を行うことも可能である[7]。

3.3 音響モデルの学習

上記で得られた音声認識結果と音声データを用いて、音響モデルの更新を行う。

ML 学習は認識結果の第一候補 (1-best) の音素列をラベルとして行う。また、MPE 学習を行うために、各発話に対しベースライン言語モデルを用いて対立仮説の単語ラティスを生成しておく。

4. 評価実験

提案手法を衆議院審議音声により評価する。

ベースライン音響モデルおよび統計的変換モデルは 2003 年、2004 年のデータを用いて学習した。これらのデータについては人手による書き起こしが存在し、予め会議録との対応づけを行っている。この音声データは 134 時間であり、テキストデータは 1.8M 単語である。

音響特徴量は 12 次元の MFCC、ΔMFCC、ΔΔMFCC、Δパワー、ΔΔパワーの計 38 次元である。デコーダは、Julius-3.5.3 を用いた。

なお、本報告では、音声データは人手によりターン毎に分割済とする。ただし、衆議院審議音声では、会議全体の音声に対する認識結果と会議録全体のアライメントを取ることにより、自動でターン分割を行うことが可能である。予備実験により、自動分割による音声を用いた場合でも、本報告と変わらない水準の結果 (ラベル精度および音響モデルの音声認識精度) が得られることを確認している。

4.1 ラベル作成実験

2006 年、2007 年の衆議院審議音声を対象に、ラベル作成の実験を行った。会議数は 26、ターン数は 5,170、データ量は 91 時間である。認識にはベースライン音響モデルを用いた。HMM の状態数は 3000、混合数は 16 であり、MPE 学習済みである。特徴量にはターン単位の CMN および CVN を適用した。Julius のサーチパラメータは、大量のデータを処理することを想定して、軽い値に設定した (RTx2 程度)。

比較のため、以下の種々の言語モデルでラベル作成実験を行った。

処理単位としては、会議全体で一つのモデルを作成する条件と、ターン毎に個別のモデルを作成する条件を比較した。手法としては、提案手法 (「会議録、話し言葉変換」) 以外に、話し言葉用ベースラインモデル (「ベースライン」)、会議録のみから作成したモデル (「会議録」)、それらを会議録に 100 倍の重みをかけて混合した biased LM(「biased LM」)、会議録モデルのポーズ位置にフィラーのエントリのみを追加したモデル (「会議録、フィラー」) のそれぞれで認識を行った。ベースラインモデルは 1999 年から 2005 年の 7 年分の会議録に話し言葉変換を適用して作成した。

音声認識により得られたラベルの精度を表 1 に示す。Corr.(単語正解率) および Acc.(単語正解精度) は人手の書き起こしを正解として算出している。

表 1 音響モデル学習用ラベル精度

	作成単位	Corr.	Acc.
ベースライン	-	82.3	79.5
会議録	会議	83.6	81.3
biased LM	会議	86.5	83.9
会議録、話し言葉変換	会議	86.3	83.7
会議録	ターン	86.1	83.5
会議録、フィラー	ターン	88.7	86.2
会議録、話し言葉変換	ターン	94.0	92.1

会議単位の条件では、biased LM や提案手法で話し言葉スタイルに対処した場合、会議録単独のモデルよりも高い精度になった。ただし、26 の会議に対し、提案手法ではコンパクトなモデルが構築できたのに対し (100MB)、biased LM ではきわめて大きなサイズとなった (1.6GB)。したがって、biased LM をターン単位の処理に適用するのは、非現実的である。

ターン単位の条件では、会議単位よりも全般に精度が高くなった。会議録のみを用いた場合は、先行研究の手法に対応する[1)4)]。また、会議録にフィラーを追加したモデルは、簡易な話し言葉向け言語モデルとなっており、話し言葉の現象のうちフィラーの挿入のみに着目

⋆1 http://julius.sourceforge.jp/

Vol.2009-SLP-77 No.8
2009/7/17

発話

総理 おっしゃっ た とおり 、 これ は 、 我が国 いー 、 に のみ ならず 、 韓国 、 周辺 国 、 うー アジア 、 あー この 地域 全体 にとって 大きな 脅威 で あり ます

会議録

総理がおっしゃったとおり、これは、我が国のみならず、韓国、周辺国、アジア、この地域全体にとって大きな脅威であります 。

作成されたラベル

総理 おっしゃっ た とおり 、 これ は 、 我が国 いー 、 のみ ならず 、 韓国 、 周辺 国 、 うー アジア 、 あー この 地域 全体 にとって 大きな 脅威 で あり ます

図 3 作成されたラベルの例

し、かつ文脈を考慮しない場合に相当する。提案手法では、会議録のみの場合よりも認識精度で 8.6 %上回り、また会議録にフィラーを追加したモデルよりも 5.9 %上回った。結果として、正解単語の 94 %を再現できた。

提案手法により作成されたラベルの例を図 3 に示す。この例では、助詞「が」の脱落や「いー」等のフィラーの挿入について正しいラベルが得られた。助詞「に」の挿入については復元できなかったが、この変換パターンはそもそも存在しなかったため、予測できなかった。

誤りの 7.9 %の内訳は、置換誤り 3.0 %、脱落誤り 3.1 %、挿入誤り 1.8 %であった。脱落誤りと挿入誤りに偏りが大きく、図 3 の例と同様、単音節の助詞やフィラーが目立った。挿入誤りでは、「え、お、あ、を、ま、い、と」の 7 語で全体の 52 %を占め、また脱落誤りでは、「え、お、あ、は、と、い、ま、を、い、で」の 10 語で全体の 50 %を占めた。置換誤りでは偏りは小さかったが、「ていう →という」、「て → と」、「え → い」、「の → あの」等、一音節の異なりの例が特に多かった。

4.2 音声認識実験

提案手法により作成したラベルを用いて音響モデルの学習データを追加し、音声認識実験による評価を行った。ベースラインモデルは、2003 年、2004 年のデータ (134 時間) を用いて人手の書き起こしラベルにより学習したものである。追加データは前節でラベルを作成した 2006 年、2007 年の 91 時間分である。会議録との比較や信頼度を用いたフィルタリングは今回行わなかった。また、比較のため、同じデータに対して人手のラベルを用いて学習を行った場合も評価する。

表 2 音声認識精度 (Word Acc.)

	学習基準	2008 年 2 月 26・29 日 予算委員会	2008 年 10 月 7 日 予算委員会
ベースライン	ML	85.4	77.6
提案手法	ML	85.9	78.4
人手ラベル (参考)	ML	85.9	78.5
ベースライン	MPE	86.8	79.2
提案手法	MPE	87.4	80.3
人手ラベル (参考)	MPE	87.6	80.4

学習は ML、MPE の二つの基準により行う。HMM の状態数は 5000、混合数は 32 であり、特徴量にはターン単位の CMN、CVN および VTLN を適用した。VTLN を行う際のラベルは、小規模なモデルを用いた高速なデコーディングにより生成した。言語モデルは、ラベル作成実験のベースライン言語モデルと同一である。

テストセットは 2008 年 2 月 26 日・29 日の予算委員会 (2.4 時間、27704 単語、121 ターン、話者数 15 名) および 2008 年 10 月 7 日の予算委員会 (3.9 時間、47382 単語、211 ターン、話者数 17 名) である。Julius のサーチパラメータは認識精度が飽和する十分大きな値に設定した (RTx10 程度)。

単語認識精度を表 2 に示す。ML 学習では、2008 年 2 月のデータで 0.5 % (有意水準 5 %で有意)、2008 年 10 月のデータで 0.8 % (有意水準 1 %で有意)、ベースラインより精度が向上し、学習データの追加による効果が確認できた。また、人手のラベルを用いた学習の場合と比較して、ほぼ同じ性能を実現できた。

MPE 学習の場合では、2008 年 2 月のデータで 0.6 % (有意水準 5 %で有意)、2008 年 10 月のデータで 1.1 % (有意水準 1 %で有意)、ベースラインより精度が向上した。識別学習はラベルの品質の影響を受けやすいと考えられたが、MPE の場合でも人手ラベルによるモデルと同等の性能となった。

5. おわりに

本稿では、統計的話し言葉変換を用いた準教師つき学習のための効果的なラベルの作成手法について提案した。提案手法を用いて、衆議院審議音声に対して、従来手法よりも高い精

Vol.2009-SLP-77 No.8
2009/7/17

度のラベルが作成可能であることを実験により確認した。さらに、作成したラベルを用いて音響モデルの学習データを追加することで、ML 学習、MPE 学習のいずれの場合でも、ベースラインモデルよりも認識精度が向上した。得られた認識精度は、人手のラベルを用いてモデルの更新を行った場合と同等の水準となった。これにより、人手のラベルが存在しない場合でも音響モデルの更新が可能であることを示した。提案手法を用いれば、国会で収録されている音声データをそのまま音響モデルの学習データとして利用できることになる。また、総選挙や内閣の改造等により話者の集合が変化した場合でも、容易に音響モデルの更新を行えるようになる。

今回、追加データがベースラインのデータよりも少量であるため、精度の向上はやや限定的な程度にとどまったが、今後さらに大量の追加データを用いた評価を行いたいと考えている。また、講義などの他のタスクへの応用についても検討したい。講義の場合、テキストデータとしては教科書やノートテイクで作成されたテキスト[8]などが利用可能と考えられる。

参 考 文 献

1) L.Lamel, J.Gauvain, and G.Adda. Investigating lightly supervised acoustic model training. In *ICASSP*, volume1, pages 477–480, 2001.
2) H.Y.Chan and P.Woodland. Improving broadcast news transcription by lightly supervised discriminative training. In *ICASSP*, volume1, pages 737–740, 2004.
3) D.Povey and P.C.Woodland. Minimum phone error and i-smoothing for improved discriminative training. In *ICASSP*, pages 105–108, 2002.
4) M.Paulik and A.Waibel. Lightly supervised acoustic model training on epps recordings. In *INTERSPEECH*, pages 224–227, 2008.
5) Y.Akita and T.Kawahara. Topic-independent speaking-style transformation of language model for spontaneous speech recognition. In *ICASSP*, volume4, pages 33–36, 2007.
6) Y.Akita and T.Kawahara. Generalized statistical modeling of pronunciation variations using variable-length phone context. In *ICASSP*, volume1, pages 689–692, 2005.
7) S.Petrik and G.Kubin. Reconstructing medical dictations from automatically recognized and non-literal transcripts with phonetic similarity matching. In *ICASSP*, volume4, pages 1125–1128, 2007.
8) 勝丸徳浩、秋田祐哉、森信介、河原達也. 大学講義のノートテイク支援のための音声認識用言語モデルの適応. In 情報処理学会研究報告, pages SLP–72–5, 2008.