Vol.2010-SLP-84 No.3
2010/12/20

国会音声認識システムの音響・言語モデルの半自動更新

秋田祐哉[†1,†2] 三村正人[†1]
Graham Neubig [†2] 河原達也[†1,†2]

音声認識システムの性能を維持するためには音響モデルと言語モデルを継続的に更新することが求められるが，このための労力とコストは小さくない．本稿では，国会音声認識システムにおいてモデル更新の負担を軽減するために検討した半自動更新の枠組みについて報告する．この枠組みは言語モデルの発話スタイル変換と音響モデルの準教師つき学習に基づいており，更新に必要な話し言葉の学習テキスト・ラベルを会議録から自動的に生成する．総選挙後の国会審議音声のためにモデル更新を行い評価したところ，音声認識精度の改善が示された．本稿ではさらに，会議録に発話スタイル変換を適用して擬似的な正解テキストを生成することによる，書き起こし不要の音声認識精度推定についても述べる．

Semi-automated Update of Automatic Transcription System for the Japanese National Congress

YUYA AKITA,[†1,†2] MASATO MIMURA,[†1]
GRAHAM NEUBIG[†2] and TATSUYA KAWAHARA [†1,†2]

Update of acoustic and language models is vital to maintain performance of automatic speech recognition (ASR) systems. To alleviate efforts for updating models, we propose a “semi-automated” framework for the ASR system of the Japanese National Congress. The framework consists of our speaking-style transformation (SST) and lightly-supervised training (LSV) approaches, which can automatically generate spoken-style training texts and labels from documents like meeting minutes. An experimental evaluation demonstrated that this update framework improved the ASR performance for the latest meeting data. We also address an estimation method of the ASR accuracy based on SST, which uses minutes as reference texts and does not require verbatim transcripts.

1. はじめに

話し言葉音声認識の研究対象は学術講演[1),2)]や大学講義[3),4)]，議会[5)]などの多様な音声に拡大してきており，我々も国会向けの音声認識システムの開発を進めている[6)]．このような音声認識システムでは，運用の間に新しい話題や話者が生じるため，これらの特徴を適当なタイミングで音響モデルや言語モデルに反映する必要がある．

運用中の音声認識システムの更新は音声対話システムにおいてしばしば実施されている．しかしこれはシステムを早急に開発・適応することが主眼である．いったん音声対話システムを構築した後は話題の変化は少なく，また音響条件が変わることもまれであると考えられるため，運用初期に集中的に行って性能の改善をすみやかに収束させるよう行われている．一方，議会の音声では話題や話者が変化し続けているほか，機材の変更による音響条件の変化も発生する．したがって音声認識システムでは音響・言語モデルを継続的に更新する必要があるが，一定量の書き起こしを更新のたびに用意することは実質的には不可能である．

国会では各会議の会議録を作成しているが，会議録では可読性の向上のために編集が行われており，たとえばフィラーは完全に削除され，また口語的な表現も丁寧な表現に置き換えられている．このため会議録を言語モデルの学習テキストや音響モデルの学習ラベルとして直接利用することはできない．これに対して，我々は言語モデルの発話スタイル変換手法をこれまでに提案している[7)]．本手法では話し言葉の N-gram エントリと統計量を会議録のような文書調のテキストから生成する．さらに我々は発話スタイル変換を音響モデルの学習用ラベル生成にも応用し，準教師つき学習を実現している[8)]．これらの手法により構築された音声認識システムの有効性は国会の審議音声における評価で実証されている．

本稿では，これらの枠組みの継続的なモデル更新への拡張について述べる．発話スタイル変換では，たとえば言語モデル補間における混合重みのような，手動によるパラメータの調整が不要である．このため手作業で編集された会議録が音声とともに提供されれば以降の処理は自動であり，「半自動」で更新作業が行えるといえる．本稿ではさらに，システムの性能を継続的に監視するための，忠実な書き起こしを必要としない音声認識精度の推定についても述べる．

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

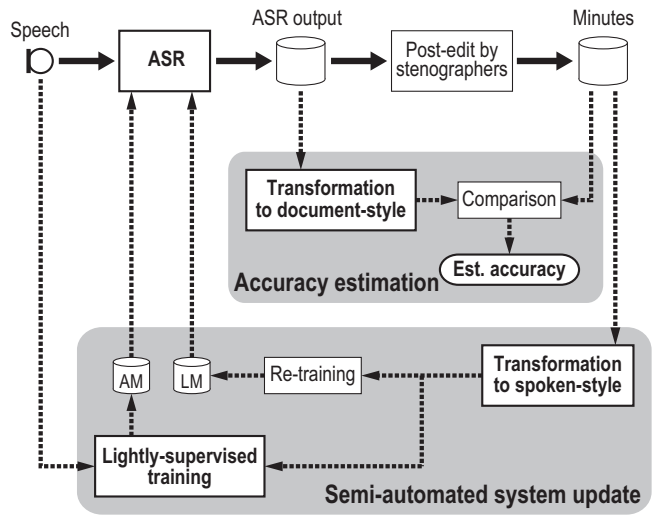

図 1 更新の枠組み

2. 半自動更新の枠組み

2.1 枠組みのあらまし

図 1 に国会（衆議院）の会議の音声認識システムを更新する枠組みを示す．各会議の審議音声は音声認識システムにより自動的に書き起こされる．この音声認識結果は，衆議院の速記者により，衆議院の厳密な表記基準に沿うように編集されて会議録として構成される．具体的には，音声認識誤りの修正のほか，フィラーなどの非流暢現象の除去や口語表現の置換などであるが，内容の要約は行われない．音声および対応する会議録はいったん蓄積されたあと，発話スタイル変換と準教師つき学習をこれらに適用して音響モデル・言語モデルの更新を行う．

国会は年間に 2〜3 回開かれ，通常国会や主な臨時国会は 2 ヶ月〜6 ヶ月ほど行われる．会期中はおおむね毎日何らかの会議が開催されるため，会期中に音声認識システムを停止させることはできない．したがって，我々は国会の閉会中にそれまでに蓄積したデータを用いて更新を行うことを想定している．

2.2 言語モデルのための発話スタイル変換

我々が提案している発話スタイルの変換手法[7] は統計的機械翻訳の枠組み[9] に基づいている．統計的機械翻訳では，元言語の文 W に対して，事後確率 $P(V|W)$ を最大とするような対象言語の文 V が出力される．$P(V|W)$ は，実際にはベイズ則を用いて式（1）のように求められる．

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

本研究では文書調のスタイルと話し言葉のスタイルを異なる言語と見なし，それぞれ W と V と記す．そして，式（1）から導かれる式（2）に基づき話し言葉の言語モデル $P(V)$ を推定する．

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

条件付き確率 $P(V|W)\cdot P(W|V)$ は変換モデルであり，忠実な書き起こしと対応する会議録からなるパラレルコーパスを用いて学習される．N-gram 言語モデルでは，この変換は実際には式（3）のように N-gram 出現頻度 N_{LM} に対して行われる．

$$N_{LM}(v) = N_{LM}(w)\frac{P(v|w)}{P(w|v)} \tag{3}$$

ここで w と v は変換元・変換先のパターンであり，$N_{LM}(w)$ と $N_{LM}(v)$ がこれらを含む N-gram エントリの頻度である．変換パターン $w\cdot v$ は文脈として前後の単語を含んでいる．また，学習データのスパースネスの問題に対処するために，文脈を品詞にバックオフしたパターンも用いる[7]．推定された N-gram エントリと頻度を用いて，話し言葉スタイルの言語モデルが構築される．

2.3 音響モデルの準教師つき学習

我々の発話スタイル変換手法は会議録から話し言葉スタイルの N-gram 統計量を求めることはできるが，音響モデルの学習に必要な書き起こしテキストそのものを復元することはできない．そこで我々の準教師つき学習手法[8] では，審議音声の話者ターン（同一話者の継続する区間）ごとに，そのターン専用の言語モデルを発話スタイル変換により構築する．これを用いて音声認識を行うことで，音響モデルの学習に利用できる高精度の書き起こしを得ることができる．

準教師つき学習の手順は次の通りである．まず，音声の各話者ターンに対して，会議録中の対応するセグメントから発話スタイル変換を用いて話し言葉スタイルの N-gram エントリ

と統計量を求め，言語モデルをそれぞれ構築する．このモデルはきわめて制約の強いモデルであるが，一方でフィラー挿入などの話し言葉特有の現象を予測することもできる．次にこのモデルを用いてターンの音声認識を行い，認識仮説を生成する．そして最良の音素仮説を音素ラベルとして採用して標準的な HMM 学習を行う．なお，最小音素誤り（MPE）基準のような識別学習では競合仮説も生成する．

2.4 システムの更新

本稿では，我々が以前に報告した音声認識システム[6] を 2009 年の第 172 回・第 173 回国会[⋆1]のために更新したものをベースラインシステムとする．言語モデルは単語 trigram モデルで，第 145 回〜第 171 回国会（1999 年〜2009 年）の衆議院会議録を学習テキストとして用いた．学習テキストのサイズは 168M 単語である．これらのテキストに対して発話スタイル変換を行い，話し言葉スタイルの言語モデルを構築した．言語モデルの語彙サイズは 64K である．音響モデルは MPE 学習による triphone HMM モデルで[10]，2003 年〜2007 年の間に収集された 225 時間の衆議院審議音声を学習データとして用いた．

衆議院では 2009 年の夏（第 171 回国会の後）に総選挙が行われた．この結果 100 名以上の議員が交代し，直後の第 172 回国会で政権も交代となり，内閣が一新された．このため，ベースラインシステムのモデルはこれらの新しい議員・閣僚を十分にカバーしていないと考えられる．

これに対して，我々は第 172 回・第 173 回のデータを用いてベースラインシステムを第 174 回国会（2010 年 1 月〜）向けに更新した．具体的には，第 173 回の会議から 95 時間の音声を収集し，我々の準教師つき手法を用いて音素ラベルを生成した．この 95 時間の音声（推定ラベル）とベースラインシステムの音響モデル学習に用いられた 225 時間分の音声（完全な書き起こし）から音響モデルを再学習した．言語モデルについては，第 172 回・第 173 回の全会議の会議録テキスト（1.3M 単語）をベースラインシステム用の学習テキスト（168M 単語）に追加し，計 170M 単語のテキストに対して発話スタイル変換を適用して新たなモデルを構築した．

2.5 評 価 実 験

我々は第 174 回の会議の審議音声を用いてシステム更新の効果の評価を行った．テストセットとして利用したのは委員会審議 3 回分で，合計の単語数は 123,405，文字数は 230,979 である．ベースラインシステムでの未知語率は 0.48%で，システム更新により新たに 310 単語が語彙に加わったものの未知語率に変化はみられなかった．デコーダとしては Julius[11] rev.4.1 を用いた．

表 1 モデル更新による単語・文字正解精度の改善

Systems	Word		Chacacter	
	Corr.	Acc.	Corr.	Acc.
Baseline (for 172nd/173rd)	82.5%	79.4%	88.0%	85.5%
AM adaptation	82.0%	78.7%	87.6%	85.0%
AM re-training	83.5%	80.5%	88.9%	86.5%
LM re-training	82.6%	79.5%	88.1%	85.6%
AM&LM re-training (for 174th)	83.6%	80.6%	89.0%	86.6%

表 1 はベースラインシステムと更新したシステムによる単語正解率・正解精度および文字正解率・正解精度である．比較のために，音響モデル更新で利用した 95 時間の音声と推定ラベルを用いてベースライン音響モデルに MLLR 適応を行ったものも評価している．この適応モデルでは性能に改善は見られず，むしろ低下が見られた．これは MPE 学習によるモデルが MLLR 適応によりゆがめられたためと推測される．一方，再学習の枠組みはモデルの一貫性が維持され，かつ追加データを十分に活用することができているといえる．再学習モデルにより，単語・文字誤りに対してそれぞれ 5.1%・6.5%（相対値）の削減が得られた．また言語モデルの更新により，単語・文字誤りの双方に対して 0.6%の削減となった．音響モデルと言語モデルの両方の更新により，単語・文字誤りの削減率は最終的に 5.9%・7.3%となった．この結果は，我々の更新の枠組みが音声認識性能の改善に効果的に機能したことを示している．

3. 会議録を用いた音声認識精度の推定

音声認識システムの運用の際，性能を監視することは不可欠である．また性能の測定は学習データの選択にも有用である．事実，国会の会議ではしばしば議論の紛糾が見られ，このような箇所は会議録から除外されることもある．紛糾した箇所は音声認識でも誤りやすいことからシステム更新には不都合であり，学習データから除いておくことが望ましい．

しかし，忠実な書き起こしが参照できないために正確な精度の算出は不可能である．そこで我々は，通常の精度測定の代わりに，会議録を用いて音声認識結果の精度を推定する手法

⋆1 第 172 回国会は内閣総理大臣を選出するための特別国会で会期は 4 日間と短いため，第 173 回国会とあわせて扱う．

を提案する．ここでは，図 1 に示したように，音声認識結果に対して発話スタイル変換[12]を行って文書調のスタイルに変換する．これを対応する会議録と比較することで精度を算出する．

3.1 会議録へのスタイル変換

文書調のスタイルへの変換手法も統計的機械翻訳に基づいている．本手法では式（1）での言語 W・V を入れ替える．

$$P(W|V) = \frac{P(V|W)P(W)}{P(V)} \tag{4}$$

したがって変換の式は次式のように改められる．

$$\hat{W} = \arg\max_{W} P(V|W)P(W) \tag{5}$$

ここで $P(V|W)$ は話し言葉スタイルから文書調のスタイルへの変換を規定する変換モデルである．また $P(W)$ は文書調スタイルの言語モデルで，変換後の単語列の妥当性を確保するためのものである．この言語モデルは会議録から学習される．

変換モデルとして，我々は文書調スタイルと話し言葉スタイルの単語履歴 $\{w\}\cdot\{v\}$ を考慮した条件付き単語確率を導入する．

$$P(V|W) \approx \prod_{i=1}^{K} P(v_i|v_1,\ldots,v_{i-1},w_1,\ldots,w_K) \tag{6}$$

式（6）における文脈長は，実際には 3 に制限されている．これらの確率はパラレルコーパスを用いて推定できる．このモデルはいわゆる「雑音のある通信路」モデルであるが，これを対数線形モデルに拡張して，結合確率や他の特徴を導入することが可能である[12]．

入力単語列 $\{w_i\}$ に対して変換仮説が生成され，重み付き有限状態トランスデューサ（WFST）に基づくデコーダにより最良の仮説が選択される．

3.2 評 価 実 験

スタイル変換モデルの学習には，158M 単語の会議録テキストと 2.8M 単語のパラレルコーパスを用いた．なお，本実験で用いたテストセットと音声認識システムは 2 節におけるものとは異なる．

テストセットには委員会の 3 会議を利用した．これらの会議には合計で 332 個の話者ターンが含まれる．書き起こしにおける単語数・文字数はそれぞれ 77,007 と 126,335 であり，会議録における単語数・文字数はそれぞれ 71,115 と 115,693 である．速記者による編集のために，会議録における単語数・文字数は書き起こしよりも 8%少ない．

表 2　推定精度と実際の精度

	Actual w/ transcripts	Estimation w/ minutes	Difference	Correlation
Word corr.	84.6%	84.9%	+0.3%	0.92
Word acc.	82.5%	80.7%	-1.8%	0.88
Char. corr.	87.2%	86.6%	-0.6%	0.88
Char. acc.	85.3%	82.6%	-2.7%	0.88

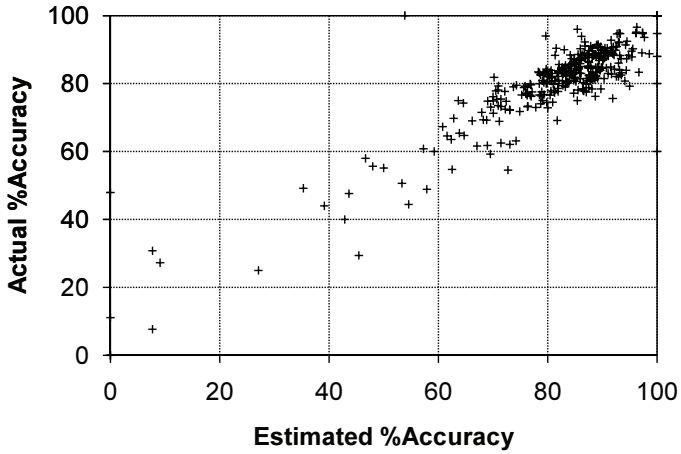

図 2　推定の文字正解精度と実際の文字正解精度の相関

単語と文字の正解率・正解精度を表 2 に示す．実際の正解率・正解精度は音声認識結果と書き起こしの比較により計算されたのに対して，推定の正解率・正解精度は変換された音声認識結果と会議録との比較により算出されている．単語・文字正解率については，発話スタイル変換に基づく推定は実際の正解率と近い値を示している．一方，単語・文字正解精度では提案手法は実際の値を下回る精度を推定した．これは，単語よりも長い単位の挿入は我々の発話スタイル変換では扱えないためである．文字正解精度に関する推定値と実際の値の差は平均で 2.7%であった．

我々は各話者ターンに対しても正解率・正解精度を求め，推定値と実際の値との間の相関係数を計算した（表 2）．また図 2 に文字正解精度における 332 ターンの相関を示す．テストセット中の 332 ターンに対して，いずれの正解率・正解精度も高い相関となった．これら

Vol.2010-SLP-84 No.3
2010/12/20

の結果は，提案法による推定精度が話者ターン単位のデータ選択に有用であることを示している．

4. おわりに

本稿では，国会音声認識システムに対する我々の半自動更新の枠組みについて述べた．音響モデル・言語モデルの構築では発話スタイル変換と準教師つき学習が用いられ，更新のための学習データが自動的に生成される．評価実験では，この更新により音声認識性能が改善することが示された．また我々は発話スタイル変換を用いた音声認識結果の精度推定も提案し，この変換に基づく会議録との比較が音声認識結果の正解率を推定できることが評価実験により確かめられた．本稿では国会の審議音声に対する枠組みを述べたが，この枠組みは講義など別のドメインに対しても適用可能である．

謝辞：本研究は JST CREST 及び科学研究費補助金によって行われた．

参考文献

1) Lamel, L., Adda, G., Bilinski, E. and Gauvain, J.: Transcribing Lectures and Seminars, *Proc. Eurospeech*, pp.1657–1660 (2005).
2) Nanjo, H. and Kawahara, T.: Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition, *IEEE Trans. Speech & Audio Proc.*, Vol.12, No.4, pp.391–400 (2004).
3) Glass, J., Hazen, T., Cyphers, S., Malioutov, I., Huynh, D. and Barzilay, R.: Recent Progress in the MIT Spoken Lecture Processing Project, *Proc. Interspeech*, pp.2553–2556 (2007).
4) Kawahara, T., Nemoto, Y. and Akita, Y.: Automatic Lecture Transcription by Exploiting Presentation Slide Information for Language Model Adaptation, *Proc. ICASSP*, pp.4929–4932 (2008).
5) Lamel, L., Gauvain, J.-L., Adda, G., Barras, C., Bilinski, E., Galibert, O., Pujol, A., Schwenk, H. and Zhu, X.: The LIMSI 2006 TC-STAR EPPS Transcription Systems, *Proc. ICASSP*, Vol.4, pp.997–1000 (2007).
6) Akita, Y., Mimura, M. and Kawahara, T.: Automatic Transcription System for Meetings of the Japanese National Congress, *Proc. Interspeech*, pp.84–87 (2009).
7) Akita, Y. and Kawahara, T.: Statistical Transformation of Language and Pronunciation Models for Spontaneous Speech Recognition, *IEEE Trans. Audio, Speech, & Language Proc.*, Vol.18, No.6, pp.1539–1549 (2010).
8) Kawahara, T., Mimura, M. and Akita, Y.: Language Model Transformation Applied to Lightly Supervised Training of Acoustic Model for Congress Meetings, *Proc. ICASSP*, pp. 3853–3856 (2009).
9) Brown, P., Pietra, S., Pietra, V. and Mercer, R.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol.19, No.2, pp.263–311 (1993).
10) Povey, D. and Woodland, P.: Minimum Phone Error and I-smoothing for Improved Discriminative Training, *Proc. ICASSP*, Vol.1, pp.105–108 (2002).
11) Lee, A. and Kawahara, T.: Recent Development of Open-Source Speech Recognition Engine Julius, *Proc. APSIPA*, pp.131–137 (2009).
12) Neubig, G., Akita, Y., Mori, S. and Kawahara, T.: Improved Statistical Models for SMT-based Speaking Style Transformation, *Proc. ICASSP*, pp.5206–5209 (2010).