議会の会議録作成のための音声認識 ー 衆議院のシステムの概要 ー

河原 達也[1,a)]

概要：衆議院で 2011 年度より運用されている音声認識を用いた会議録作成システムについて述べる。このシステムでは、原則すべての本会議・委員会の審議において、発言者のマイクから収録される音声を音声認識システムで書き起こし、会議録の原稿作成支援を行う。議会審議のような話し言葉音声に対して高い認識性能を実現するために、審議音声と会議録を “そのまま” 効果的に活用する枠組みを考案・実装した。これは、発言の忠実な書き起こしと編集された会議録テキストとの間の統計的な言語モデル変換に基づいている。この枠組みにより、忠実な書き起こしを作成することなく、音響モデルと言語モデルの学習を可能にし、長期的なモデルの半自動更新も可能にしている。運用後も含めて、12 年分の会議録テキストと約 1000 時間の審議音声で音響・言語モデルを構築することにより、約 90%の文字正解率を実現した。本稿では、音声認識結果を編集するソフトウェア (エディタ) やシステムの運用に関しても述べる。

キーワード：音声認識，会議録，議会，衆議院，音響モデル，言語モデル，準教師付き学習

Overview of Automatic Speech Recognition for Transcription System in the Japanese Parliament (Diet)

Tatsuya Kawahara[1,a)]

Abstract: This article describes a new automatic transcription system in the Japanese Parliament which deploys our automatic speech recognition (ASR) technology and has been in official operation since April 2011. The speaker-independent ASR system handles all plenary sessions and committee meetings to generate an initial draft, which is corrected by Parliamentary reporters. To achieve high recognition performance in spontaneous meeting speech, we have investigated an efficient training scheme with minimal supervision which can exploit a huge amount of real data. Specifically, we have proposed a scheme of statistical language model transformation, which fills the gap between faithful transcripts of spoken utterances and final texts for documentation. Once this mapping is trained, we no longer need faithful transcripts for training both acoustic and language models. The scheme also realizes a sustainable ASR system which evolves, i.e. update/re-train the models, only with speech and text generated during the system operation. After its initial deployment in 2010, the system has been improved with accumulated data of 1000-hour speech, consistently achieving character correctness of approximately 90%.

Keywords: Speech recognition, Transcription, Parliament, House of Representatives, Acoustic model, Language model, Lightly supervised training

[1] 京都大学
Kyoto University, Kyoto 606-8501, Japan
[a)] http://www.ar.media.kyoto-u.ac.jp/diet/

1. はじめに

明治 23 年 (1890 年) に我が国に議会が開設されて以来百年以上にわたり、会議録の作成は手書き速記によって行われてきた。今世紀になって衆参両院において、速記者の新規採用・養成が停止され、新たな会議録作成方法が模索された。[*1]様々な検討をふまえて、衆議院において音声認識技術を用いたシステムが導入された。

このシステムでは、原則すべての本会議・委員会の審議において、発言者のマイクから収録される音声を音声認識システムで書き起こし、会議録の草稿を生成する [1], [2], [3]。ただし、音声認識には一定の誤りが不可避な上、話し言葉の発言を忠実に書き起こしても会議録とはならないため、高度な編集スキルを有する速記者・校閲者の役割がなくなるわけではない。

2. システムの基本的要件

会議録作成を目的とする本システムに求められた第一の要件は、高い認識精度である。「認識精度が少々低くても一から全部入力するよりはよいだろう」といった見方もあるが、専門的に編集作業する人にとっては、認識誤りがある程度以上多くなると、不快であるばかりか、思考や作業の妨げとなる。予備調査の結果、本システムでは文字認識精度 85%が要件とされた。[*2]ほぼ原稿の読上げの形式で行われる本会議の審議音声に対して、90%の認識精度を実現するのは容易であるが、衆議院では大半の審議が委員会で行われ、原稿の読上げでない丁々発止の議論が行われる。非常に自発性の高い音声である。欧州議会の音声翻訳を目指して研究開発が行われた TC-STAR プロジェクト [4], [5] では本会議のみを扱っていたが、我々の調査では、本会議と委員会でフィラー率や認識精度で大きな違いがあることが示されている。

第二の要件は、速い処理速度である。衆議院では、通常 5 分の作業単位に区切って会議録作成の作業を行う。すなわち当該会議の最中でも、最初の作業単位から順次すみやかに作業に着手できる必要がある。したがって、完全なリアルタイム処理を保証する必要はないが、速いターンアラウンドが要求される。1 日に複数の会議が朝から晩まで行われることと、計算資源が限られることを考慮すると、事実上音声認識処理にかかる時間の実時間比 (Real Time Factor: RTF) を 1 に近くする必要がある。

第三の要件として、衆議院では『用字例』に則って、表記・文字遣いが厳格に定められているので、音声認識システムの語彙・単語辞書もこれに従う必要がある。新聞記事や Web などの他の言語資源を利用することは、表記の揺れにつながるので、事実上できない。衆議院の過去の会議録テキストのみを用いて、単語辞書・言語モデルを作成する必要がある。

『日本語話し言葉コーパス』(CSJ) を含む、従来の話し言葉を対象とした音声認識システムでは、これらの要件を満たすことは到底できない。

3. 基本的アプローチ –言語モデル変換

3.1 発言の忠実な書き起こしと会議録テキストとの差異

現代の音声認識システムの音響モデル（＝音素毎の周波数パターンのモデル）と言語モデル（＝単語系列の頻度パターンのモデル）は、大規模な統計的モデルに基づいているので、大規模な学習コーパス（＝音声とテキストのデータベース）が鍵となっている。本システムの研究開発においても、まず必要となったのが大規模なコーパスである。

幸い、国会には審議音声と会議録テキストの大規模なアーカイブが存在する。年間の審議音声は 1000 時間以上にも及び、その会議録テキストは 1200 万単語以上になる。1999 年 (第 145 回国会) 以降の会議録テキストは電子化されており、これは新聞記事にも匹敵するスケールである。したがって、音声と会議録テキストを大規模に収集することは容易である。

しかしながら、公式の会議録は速記者や校閲者の編集過程を経て、実際の発言と異なる部分がある。これにはいくつかの理由がある [6], [7]。話し言葉と書き言葉の違い、フィラー (「えー」「あのー」など) や言い直しなどの言い淀み、句末・文末などの談話的に冗長な表現 (「〜ですね」など)、文法的訂正などである。[*3]

以上の理由から、言い淀みなども含めて実際の発言を忠実に書き起こしたコーパスを構築する必要がある。我々が当初構築したこのようなコーパスは、音声で 225 時間、テキストで約 270 万語の規模である。これは、公式の会議録との対応付けも行っている。このようなコーパスは、満足のいく性能を得るために必要不可欠であるが、膨大なコストと時間を要し、現実的には大規模な審議データのごく一部にしか作成できない。ある程度の規模の書き起こし付きコーパスを用意できたとしても、会議室の音響環境や話者集合、話題・語彙は年々変化していくので、更新していく必要がある。

国会審議の大規模なアーカイブをより効果的に活用するために、発言の忠実な書き起こしと公式の会議録とを対応づけて、両者の違いを統計的に調査した。平均 13%の単語

[*1] 諸外国でもこの数十年の間に手書き速記から別の方式への移行が行われている。大半が、録音されたものをワープロソフトで書き起こす (テープ起こし) 方式であるが、イタリア議会では録音されたものをリスピークする形でディクテーションソフトが用いられている。

[*2] 速記者の一般的な意見では、90%以上が望まれている。予備調査においては、80%以下だと「利用したくない」という意見が多かった。大半の審議音声区間に対して 80%以上を確保するためには、平均 85%という目標設定が妥当である。

[*3] TC-STAR のコーパス (欧州議会の英語) と比較すると、日本語では冗長性や言い淀みなどが多いが、文法的訂正は少ない。

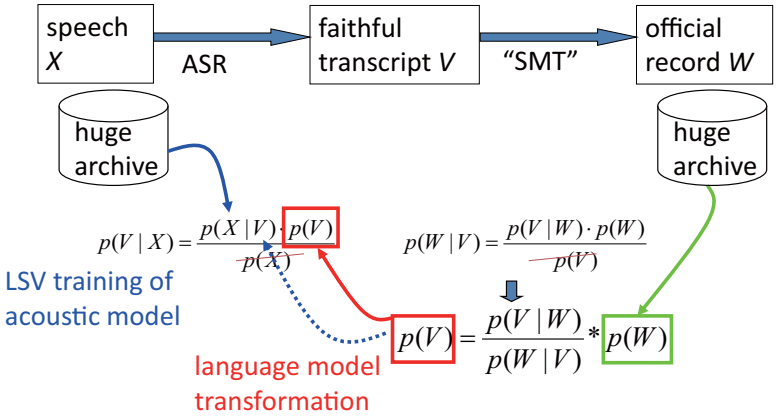

図 1　統計的言語モデル変換に基づく言語モデルと音響モデルの学習の枠組み

で差異がみられ、その半数近くはフィラーによるものであるが、フィラー以外の冗長語の削除や、口語的表現の修正 (置換)、助詞の補完 (挿入) などもかなり多い。ただし、これらの大半には規則性があり、前後の単語 (文脈) を考慮することで、統計的な枠組みでモデル化することができる。実際に、両者の差異の 93%は 1〜2 単語の削除・置換・挿入のいずれかであった。

3.2　統計的言語モデル変換による発言体の言語モデル推定

我々は、会議録のテキストから発言内容を確率的に予測する枠組み (言語モデル変換) を考案した [7], [8], [9]。これは、テキスト自体を変換するのではなく、言語モデルの統計量を変換するものである。この枠組みを図 1 に示す。

この枠組みでは、発言の忠実な書き起こし (V) と会議録テキスト (W) を別の言語とみなして、統計的機械翻訳を適用する。発言体から会議録への整形 ($V \rightarrow W$) と会議録から発言体の復元 ($V \leftarrow W$) の両方向について、各々以下のベイズ則に基づいて定式化することができる。

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

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

発言体から会議録への整形 ($V \rightarrow W$) に関しては、現代の統計的機械翻訳と同様に、上記のような単純なモデル (式 (1)) でなく、種々の素性を導入した対数線形モデルへの拡張も検討し、効果を確認している [10]。

これに対して、会議録から発言体の復元 ($V \leftarrow W$; 式 (2)) は、発言体から会議録への整形 ($V \rightarrow W$; 式 (1)) と比較して、任意性が高く、実現は容易でない。例えば、フィラーは整形の過程ではすべて削除されるが、復元の過程では任意の場所にランダムに挿入可能である。したがって、発言体の書き起こし (V) そのものを復元するのではなくて、その統計的な予測を行う言語モデル $p(V)$ を推定するのが望ましい。式 (1) を変形することにより、以下が得られる。

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

この枠組みのポイントは、発言体の書き起こし (V) に比べて、会議録テキスト (W) の量が膨大であることである。上式によって、その統計量 (言語モデル) をフルに活用して、音声認識システムに必要な発言体の言語モデル $p(V)$ を推定することができる (図 1 の右半分)。

この変換は、実際には下記のように N-gram エントリに対して行われる。本システムでは、3-gram までを考慮している。

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

ここで、v と w は個々の変換パターンであり、$w \rightarrow v$ の置換、w の削除、v の挿入の 3 通りを文脈 (前後の単語) を含めて考慮している。$N_{gram}(w_1^n)$ は、これを含む N-gram エントリの計数であり、上式によって、$N_{gram}(v_1^n)$ に補正される。翻訳モデルに相当する条件付き確率 $p(v|w)$ 及び $p(w|v)$ の推定には、発言の忠実な書き起こし V と会議録テキスト W を対応づけたパラレルコーパス (前記の学習コーパス) を利用する。単語の対応付けから条件付き確率を推定する際に、前後の単語も考慮する。例えば、フィラー「えー」の挿入は、$\{w = (w_{-1}, w_{+1}) \rightarrow v = (w_{-1}, \text{えー}, w_{+1})\}$, のようにモデル化され、この挿入によって影響を受ける N-gram エントリの計数が補正される。文脈については、品詞の情報を用いたスムージングも導入している。条件付き確率のパラメータの数は多くないので、学習コーパスのサイズは少量 (数十時間規模) でも十分である [9]。

この枠組みにより、膨大な会議録のテキストのみから、発言体の言語モデルを推定することができる。これは、コーパス混合の手法と比較して、混合重みを調整する必要がないという利点がある。また、単語辞書・言語モデルは基本的に会議録のみから構築しているので、衆議院の『用字例』を忠実に反映した語彙・表記となることが保証される。これらの点は、継続的にモデルを更新していく本システムにおいてはきわめて重要である。

3.3 音響モデルの準教師付き学習

また、この言語モデル変換の枠組みを応用して、音響モデルの準教師付き学習 (Lightly SuperVised training: LSV) を行う手法を考案した [11], [12]。国会審議には膨大な音声アーカイブがあるが、発言の忠実な書き起こしはなく、会議録テキストのみが対応づけられる。前述の通り、会議録テキストから発言の書き起こしを復元するのは容易でないが、会議録のテキストから発言内容を予測する言語モデルを推定し、これを用いて音声認識することで、書き起こしを高い精度で生成する (図 1 の左半分)。

会議のターン (10 秒〜3 分程度；平均約 1 分) 毎に、対応する会議録テキストから N-gram を求める。ここでターンを単位としたのは、会議全体だと話題や話者が多岐にわたるためである。この N-gram を式 (4) に従って、発言体のものに変換することによって、フィラーの挿入等も考慮した言語モデルが得られる。この言語モデルを用いて当該ターンの音声を認識することで、発言の書き起こしを生成する。この言語モデルは当該ターンに特化している上に、話し言葉の現象も考慮している。さらに、従来の準教師付き学習で一般的に用いられている重み付き混合モデル [13][14] と比べても非常にコンパクトである。

本手法により、忠実な書き起こしを用意する場合と同等の精度の音響モデルを学習できることが示された [12]。なお、最尤推定 (ML) だけなら、上記で最尤の音声認識結果を求めれば十分であるが、MPE などの識別学習を行うためには、ベースライン言語モデルを用いて競合仮説を生成する必要がある。

ベースラインの音響モデルを学習するために、ある程度の書き起こし付き音声コーパス (言語モデル変換のモデル学習用のパラレルコーパスと同一) は必要とするが、この枠組みは、審議音声と会議録テキストのみで、半自動的に音響モデルと言語モデルの追加学習・更新を可能にするものである。今後さらに多くのデータが蓄積されれば一層の性能向上が期待できる。また、総選挙や内閣改造に伴って議員や閣僚が交代したり、年をおって話題・語彙が変化しても、それらを反映することができる。

4. 音声認識システムの構成

音声認識システムは、上記のように構築された音響モデル・言語モデル・単語辞書 (基本的に京都大学で開発) を、有限状態トランスデューサ (WFST) に基づくデコーダ [15](NTT で開発) に統合することで構成されている。また、チャネル選択や話者区分化などの前処理も導入した。

入力音声は質問者と答弁者 (＋議長) のマイクから収録され、2 チャネルでデジタル化・入力される。会議室内の拡声設備による回り込みも多いので、周波数特徴に基づいてチャネル選択を行った上で、話者ターンへの区分化を行う [16]。このターン単位に対して、CMN と CVN 及び VTLN を適用する。音響特徴量は、12 次元の MFCC, ΔMFCC, ΔΔMFCC, ΔPower, ΔΔPower の計 38 次元で構成される。

音響モデルは状態共有トライフォン HMM(状態数 3000・混合数 16) である。モデルは MPE 基準により学習した [17]。学習データは、本システムの開発を行った 2009 年度当初では 225 時間の書き起こし付きコーパスのみであった [18], [19] が、その後前章で述べた枠組みで、順次書き起こしのないデータを追加し、2011 年末時点では約 1000 時間に及んでいる。

言語モデルは、前章で述べた通り、会議録テキストから学習したモデルを変換することによって生成される単語 3-gram モデルである。形態素解析には、NTT で開発された JTAG を用いている。語彙サイズは約 64K である。学習データは、1999 年 (第 145 回国会) 以降の会議録テキストのすべてであり、約 2 億形態素に及ぶ [20]。なお、会議録テキストに存在しないフィラーにはアノテーションがされ、認識後に自動除去することもできる。

デコーダは、高速 on-the-fly 合成を用いる WFST[15] に基づくものである。ユーザが新しい単語をいつでも登録できるように対応している [21]。

なお実際のシステムでは、審議音声は作業単位 (原則 5 分) 毎に機械的に区分化して、音声認識システムに入力される。発話の途中で区切られる場合があるので、前後に重複区間 (1 分) を設定している。

5. 音声認識システムの評価

新会議録作成システムは、2010 年 3 月に衆議院に納入され、2010 年度に試行が行われた。音声認識結果の評価は、最終的に作成された会議録テキストと照合した文字正解率 (Character Correct) により行っている。これは、通常の音声認識精度の計測法とは異なるが、文字正解率については一定の目安になることがわかっている [22]。これにより、忠実な書き起こしを作成することなく、大規模かつ継続的に性能評価を行うことが可能になっている。なお、音声認識結果に存在するフィラーは除去している。[*4]

2010 年 8 月〜11 月に行われた 60 の会議において、認識性能を評価したところ、文字正解率は平均 89.3%であった。本会議に限ると概ね 95%を達成しており、全委員会に対して 85%を下回った会議は皆無であった。処理速度に関しては、RTF で 0.5 程度であった。なお一部の会議について、忠実な書き起こしを用意して、文字認識精度 (Character Accuracy) を計測し、平均 85%を実現していることを確認している。2010 年末に音響・言語モデルの更新を行い、同

[*4] システムの目標が会議録の作成であるので、会議録と照合する方が合致している。なお、単語でなく文字を単位としているのは、形態素解析システムに依存しないことと、速記業界の評価尺度に準拠するためである。

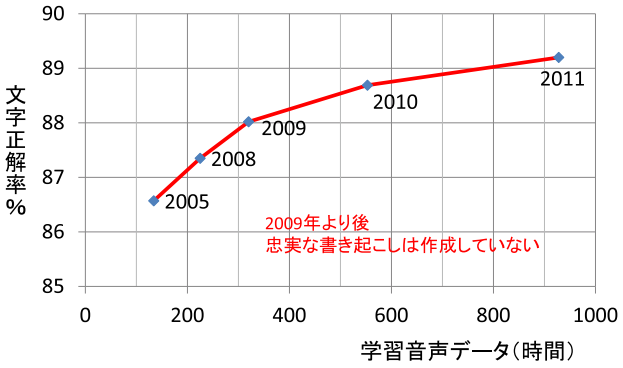

図 2 音響モデル学習データ量と文字正解率の関係 (12 会議による評価)

一のテストセットに対して、絶対値で 0.7%の改善が得られた [23], [24]。

この試行を経て、2011 年度 (2011 年 4 月) から本システムは正式に運用されている。2011 年 1 月〜12 月に行われた 118 の会議における文字正解率は平均 89.8%であった。2011 年末にも音響・言語モデルの更新を行い、絶対値で 0.5%の改善が得られた。その結果 2012 年は、ほぼすべての会議で 88%以上、平均で 90%を上回る水準になっている。

特に、この数年の間に内閣が頻繁に代わり、2011 年 3 月以降は、東日本大震災・原発事故関連の話題にシフトするという状況で、安定した認識率を維持できたのは、音響・言語モデルに一定の普遍性があることを示唆するものである。また、システムの運用・データの蓄積に伴って、継続的な性能の改善が得られているのは、提案する枠組みの有効性を示すものである。音響モデルの学習データ量と音声認識精度の関係を図 2 に示す。この文字正解率は、同一のテストセット (2011 年の 12 会議) に対して会議録と照合したものである。

ただし、フィラーを除く挿入誤り (会議録との比較であり実際に発言されたものを多く含む) は 2011 年全体の平均で 8.2%であった。その多くは、「ですね」などの冗長語であると考えられる。この除去を含む他の編集の自動化についても研究を進めている [10] が、実装については今後の課題である。

6. システムのユーザビリティと運用

上記の通り、性能要件を上回る認識精度を実現したが、認識誤りが 10%程度存在するのも事実であり、これ以外に冗長語の削除や口語表現の修正が必要な箇所も 10%弱ある。したがって、原稿作成者の役割・負担も依然大きいといえる。[*5]なお、衆議院におけるシステムの運用については、[25] も参照されたい。

[*5] 現在、原稿作成には速記者が従事しているが、今後は速記者以外の方が順次採用され、訓練を経て、従事していくことになる。

6.1 編集ソフトウェア (エディタ)

原稿作成のために、音声認識結果を修正・編集するソフトウェア (エディタ) の役割は非常に大きい。原稿作成者が会議録としてふさわしい文章の編集に集中できるよう、ラインエディタではなく、ワープロソフトのようなスクリーンエディタが採用された。エディタについては、技術者でなく速記者が設計したことに留意されたい。エディタは、審議音声と映像に時刻・発言・文字単位で簡単にアクセスすることができ、音声再生の速度を速くしたり遅くしたりすることもできる。音声認識の副次的な効果として、すべての書き起こしと音声・映像がデジタル化され、発言者や発言ごとに対応づけされることが挙げられる。

6.2 モデルの更新

システムメンテナンスのため、継続的に認識精度をモニタしており、音声認識のモデルも更新している。特に、言語モデルは新語や新しい話題を取り入れるために年に一度、国会会期の合間に更新している。ただし、新語はいつでも、ワープロソフトの単語登録機能と同様に、追加することができる。音響モデルは、内閣の大幅な改造もしくは総選挙による議員交代の際に更新されることになっている。

なお 2011 年度には、常用漢字表の改定に伴って、『用字例』の大規模な改訂補正が行われた。単語辞書・言語モデルの表記の一貫性を保証するために、学習に用いている過去の会議録テキストのデータベースについても新たな『用字例』に則するように修正を行った。これには多くの作業を要したが、数十年に一度のことと思われる。

7. おわりに

衆議院において導入された新しい会議録作成システムにおける音声認識システムについて、その概要を述べた。このシステムは、人間どうしの自然な話し言葉の音声認識としては最高水準のものと考えられる。また、システムを運用していく上で自然に蓄積される音声と会議録テキストを用いて、半自動的にモデルの追加学習・更新ができる枠組みとなっており、今後も継続的な性能改善が期待できる。

音声認識技術を用いた会議録・記録作成は、地方議会や裁判所などでも導入が進んでいる。ただし、一般の講演会や学校の講義・授業などへの展開にはまだまだ課題がある。聴覚障害者や外国人のための情報保障のために、様々な音声メディアに対して字幕を付与することが求められており、今後は、このような目的に向けて研究開発を進めていく予定である。

謝辞

本研究開発は、約 10 年にわたって行ってきたもので、数多くの協力・支援によるものである。特に、3 章で述べた方法論を実装した秋田祐哉助教及び三村正人研究員の貢献は多大である。衆議院の会議録作成システムの開発には、政瀧浩和、高橋敏、小橋川哲、堀貴明各氏をはじめとして、NTT グループの多くの方が従事された。ここに感謝の意を表したい。

参考文献

[1] 河原達也. 日本の国会における音声認識技術を用いた会議録システム (Intersteno 2009 講演要旨). 日本の速記, No. 852, pp. 12–17, (11 月号) 2009.

[2] T.Kawahara. Transcription system using automatic speech recognition for the Japanese Parliament (Diet). In *Proc. AAAI/IAAI*, pp. 2224–2228, 2012.

[3] 河原達也, 秋田祐哉, 三村正人, 政瀧浩和, 高橋敏. 衆議院会議録作成における音声認識システム — 全体の構成と評価 —. 日本音響学会研究発表会講演論文集, 3-5-5, 春季 2011.

[4] C.Gollan, M.Bisani, S.Kanthak, R.Schluter, and H.Ney. Cross domain automatic transcription on the TC-STAR EPPS corpus. In *Proc. IEEE-ICASSP*, Vol. 1, pp. 825–828, 2005.

[5] B.Ramabhadran, O.Siohan, L.Mangu, G.Zweig, M.Westphal, H.Schulz, and A.Soneiro. The IBM 2006 speech transcription system for European parliamentary speeches. In *Proc. INTERSPEECH*, pp. 1225–1228, 2006.

[6] 河原達也. 〔招待論文〕 筆記録作成のための話し言葉処理技術. 電子情報通信学会技術研究報告, SP2006-120, NLC2006-64 (SLP-64-36), 2006.

[7] T.Kawahara. Automatic transcription of parliamentary meetings and classroom lectures – a sustainable approach and real system evaluations –. In *Proc. Int'l Sympo. Chinese Spoken Language Processing (ISCSLP)*, pp. 1–6 (keynote speech), 2010.

[8] 秋田祐哉, 河原達也. 統計的機械翻訳の枠組みに基づく言語モデルの話し言葉スタイルへの変換. 電子情報通信学会技術研究報告, SP2005-108, NLC2005-75 (SLP-59-19), 2005.

[9] Y.Akita and T.Kawahara. Statistical transformation of language and pronunciation models for spontaneous speech recognition. *IEEE Trans. Audio, Speech & Language Process.*, Vol. 18, No. 6, pp. 1539–1549, 2010.

[10] G.Neubig, Y.Akita, S.Mori, and T.Kawahara. A monotonic statistical machine translation approach to speaking style transformation. *Computer Speech and Language*, Vol. 26, No. 5, pp. 349–370, 2012.

[11] T.Kawahara, M.Mimura, and Y.Akita. Language model transformation applied to lightly supervised training of acoustic model for congress meetings. In *Proc. IEEE-ICASSP*, pp. 3853–3856, 2009.

[12] 三村正人, 秋田祐哉, 河原達也. 統計的言語モデル変換を用いた音響モデルの準教師付き学習. 電子情報通信学会論文誌, Vol. J94-D, No. 2, pp. 460–468, 2011.

[13] L.Lamel, J.Gauvain, and G.Adda. Investigating lightly supervised acoustic model training. In *Proc. IEEE-ICASSP*, Vol. 1, pp. 477–480, 2001.

[14] M.Paulik and A.Waibel. Lightly supervised acoustic model training EPPS recordings. In *Proc. INTERSPEECH*, pp. 224–227, 2008.

[15] T.Hori, C.Hori, Y.Minami, and A.Nakamura. Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition. *IEEE Trans. Audio, Speech & Language Process.*, Vol. 15, No. 4, pp. 1352–1365, 2007.

[16] 小橋川哲, 浅見太一, 山口義和, 阪内澄宇, 小川厚徳, 政瀧浩和, 高橋敏, 河原達也. 衆議院会議録作成における音声認識システム — 事前音響処理 —. 日本音響学会研究発表会講演論文集, 3-5-9, 春季 2011.

[17] 三村正人, 秋田祐哉, 河原達也. 衆議院会議録作成における音声認識システム — 音響モデル —. 日本音響学会研究発表会講演論文集, 3-5-7, 春季 2011.

[18] 秋田祐哉, 三村正人, 河原達也. 会議録作成支援のための国会審議の音声認識システム. 電子情報通信学会論文誌, Vol. J93-D, No. 9, pp. 1736–1744, 2010.

[19] Y.Akita, M.Mimura, and T.Kawahara. Automatic transcription system for meetings of the Japanese national congress. In *Proc. INTERSPEECH*, pp. 84–87, 2009.

[20] 秋田祐哉, 河原達也, 政瀧浩和. 衆議院会議録作成における音声認識システム — 言語モデル —. 日本音響学会研究発表会講演論文集, 3-5-6, 春季 2011.

[21] 堀貴明, 中村篤, 山口義和, 小橋川哲, 浅見太一, 政瀧浩和, 高橋敏, 河原達也. 衆議院会議録作成における音声認識システム — 探索技術 —. 日本音響学会研究発表会講演論文集, 3-5-8, 春季 2011.

[22] 秋田祐哉, 河原達也. 国会音声における認識文と整形過程の分析. 日本音響学会研究発表会講演論文集, 2-1-6, 秋季 2009.

[23] 秋田祐哉, 三村正人, Graham Neubig, 河原達也. 国会音声認識システムの音響・言語モデルの半自動更新. 情報処理学会研究報告, SLP-84-3, 2010.

[24] Y.Akita, M.Mimura, G.Neubig, and T.Kawahara. Semi-automated update of automatic transcription system for the Japanese national congress. In *Proc. INTERSPEECH*, pp. 338–341, 2010.

[25] 猿谷豊. 衆議院における音声認識を利用した会議録作成業務. 情報管理, Vol. 66, No. 6, pp. 392–399, 2012.