

会議録作成支援のための国会審議の音声認識システム

秋田 祐哉[†] 三村 正人[†] 河原 達也[†]

[†] 京都大学 学術情報メディアセンター
〒 606-8501 京都市左京区吉田本町

あらまし 我々は国会審議の会議録作成支援を想定した音声認識システムの研究に取り組んでいる。会議録では原則として発話をすべて書き起こして記録することから、音声認識を活用する際は高い認識精度が求められる。このため、本研究では衆議院の審議音声からなるコーパスの整備を進めるとともに、これを用いたより高性能な音響モデル・言語モデル・発音辞書の検討を行っている。音響モデルについては、種々の正規化手法に加えて最小音素誤り (MPE) 学習を導入した。また言語モデルと発音辞書に関しては、話し言葉音声向けのモデルを生成するために発話スタイルの統計的変換手法を適用し、4-gram 統計言語モデルと発音の変異形を含む辞書を構築した。これらのモデルに基づく音声認識システムについて実際の会議音声における評価を行ったところ、85%の単語正解精度が得られた。

キーワード 会議, 話し言葉音声認識, 言語モデル, 音響モデル, 発音辞書

An automatic transcription system for creation of meeting records in the Japanese Congress

Yuya AKITA[†], Masato MIMURA[†], and Tatsuya KAWAHARA[†]

[†] Academic Center for Computing and Media Studies, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract We have been developing an automatic speech recognition (ASR) system for creation of meeting records of the National Congress (Diet) of Japan. Since almost all utterances in the meetings must be transcribed and recorded, an ASR system should provide accurate transcripts as drafts. Therefore, we have compiled meeting speech of the Congress into a corpus for development of an ASR system, and investigated more powerful acoustic, language and pronunciation models base on the corpus. As for acoustic model, we introduce minimum phone error (MPE) training, as well as several normalization methods. To build language and pronunciation models, our speaking-style transformation approach is applied to the corpus, and models for spontaneous speech recognition are generated. The proposed ASR system using these models was evaluated on real meeting speech of the Congress, and realized 85% of word accuracy.

Key words Meeting, Spontaneous speech recognition, Language model, Acoustic model, Pronunciation model

1. はじめに

現在、講義や講演、会議などの話し言葉音声 (Spontaneous speech) を対象とした音声認識の研究・開発が広く行われている。たとえば講義の音声認識は国内外の大学を中心に組みが進められているほか [1]~[3], 講演についても「日本語話し言葉コーパス」(CSJ) [4] を対象とした研究 [5], [6] などが報告されている。また、米国の Rich Transcription プロジェクト [7] や欧州の AMI/AMIDA プロジェクト [8] では、多人数の会議における音声認識の研究が進められている。一方、法廷や議会などの公の場を対象とした話し言葉音声認識も関心を集めており、法廷における音声認識については文献 [9] などで報告されている。議会については、代表的な研究プロジェクトとして欧州議会 (European Parliament) を対象とした TC-STAR があり、複数の公用語を持つ欧州議会における音声認識と機械翻訳の研究が多く、研究機関によって行われた [10]~[12]。また我々もこれまで衆議院の会議を対象とした音声認識の研究に取り組んでいる [13], [14]。

講義や会議の音声認識は、アーカイブ化やインデキシング・情報検索、あるいは字幕・ノートテイキングを主な目的として研究が続けられてきた。このようなアプリケーションでは音声認識により発話の内容が把握・抽出できればよく、必ずしも音声を忠実に書き起こす必要はない。これに対して法廷や議会のタスクでは発言記録の正確さがきわめて重要である。音声認識の利用に際して、インデキシングや検索などで求められる水準よりもはるかに高い認識精度が必要とされる点がこのようなタスクの特徴といえる。議会では講義や会議よりも丁寧な発話をする傾向はあるものの、発音の怠けやフィルラーなどの非流暢現象、また話者個人性など、講義や会議と同様に話し言葉特有の音響的・言語的な多様性がみられ、これが音声認識を難しくする要因となっている。高い認識精度を達成するためには、これらを効果的にカバーすることが求められる。

このような観点から、我々は国会向けの音声認識用モデルの改良や精緻化を検討し、認識システムの改善を続けている。音響モデルに関しては、識別学習を導入してより効果的なモデリングを図っている。また言語モデルや発音モデル (辞書) については、話し言葉の統計的分析に基づくスタイル変換法を適用し、効率的にモデルを生成している。本稿では、これらのモデルの構成法を性能の評価とともに述べる。以下では、まず 2. にて議会の音声認識に求められる要件を考察し、本研究の目標とするところについて述べる。次に 3. で提案システムの各モデルについて詳述し、4. にて実際の国会音声によるシステムの評価を行う。

2. 議会の音声認識

議会の音声認識は TC-STAR プロジェクトによるものがよく知られている。TC-STAR では日本の国会の本会議に相当する

European Parliament Plenary Speech (EPPS) での演説を対象としているが、このような演説は特定の人物が事前に用意した原稿に沿って比較的長い時間行うことが一般的である。このため発声も言語表現も一様で明瞭となりやすいことから、TC-STAR に参加した研究機関では 90% を超える高い単語正解精度を達成している。これに対して、我々が扱っている日本の国会は「委員会中心主義」であり、本会議よりも各委員会における議論が審議の中心となっている。したがって音声認識システムも委員会審議を主な対象とすることになるが、これには (1) 委員会ごとに話題が大きく異なる専門的な議論、(2) 閣僚・官僚・議員など、入れ替わりに発言する多数の話者、(3) 演説と比較してより自発的な発話スタイル、(4) 質問者が興奮することなどによる感情的な音声、などの特徴がみられる。音響・言語的な多様性は本会議や EPPS よりも明らかに大きく、難しい認識タスクとなっているといえる。

これらは音声認識そのものの課題であるが、一方会議録作成支援における実運用の観点からみると、音声認識システムにはさらなる課題が想定される。音声認識により会議録作成に携わる作業 (速記者) の作業効率を改善することが導入の目的であるから、必然的に備えるべき要件として、(A) 認識結果を修正する方が文をタイプし直すよりも速いと考えられる高い認識精度、(B) 議会で定められている厳格な文字・単語表記ルールに則った出力、(C) 速記者による修正を容易にするための、句読点挿入・自動整形などの後処理、(D) 会議録の作成期限に合う範囲の認識処理時間、などが考えられる。

これらの分析をもとに、本研究ではまず (A) の認識精度の目標を単語正解精度 85%・文字正解精度 90% と設定した。また (D) の認識処理時間について、これまでの研究では長い時間をかけてきたものを、より現実的な範囲に改めることとした。この設定のもとで (1) ~ (4) や (B) に挙げた問題に対処できる音響モデル・言語モデルの検討を行うこととなる。なお、我々は (C) の音声認識の後処理についても研究を進めているが [15], [16]、本稿では扱わない。

3. 国会審議のための音声認識システム

我々が取り組みを進めている音声認識システムの概要を図 1 に示す。デコーダには Julius rev.4.1 を利用し、音声認識に利用する音響モデル・言語モデル・発音辞書は、我々が構築している「衆議院審議コーパス」の音声・テキストデータをもとに学習される。本節ではこのコーパスと各モデルについて述べる。

3.1 衆議院審議コーパス

本研究ではこれまで衆議院の会議音声からなるコーパスを構築してきた。会議の種類としては予算委員会が比較的多くなっているが、外交・法務・農林水産・文部科学・国土交通などの主要委員会を中心に多くの種類の会議から構成されており、多様な話題や多くの話者を含んでいると考えられる。なお、衆議院には多くの会議室があり、それぞれの会議室ごとに部屋のサイ

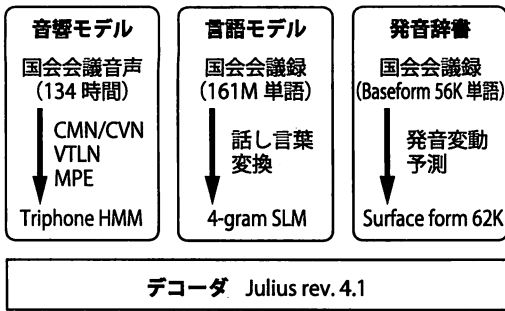


図1 国会音声認識システムの構成

Fig. 1 Components of the proposed ASR system for the National Congress

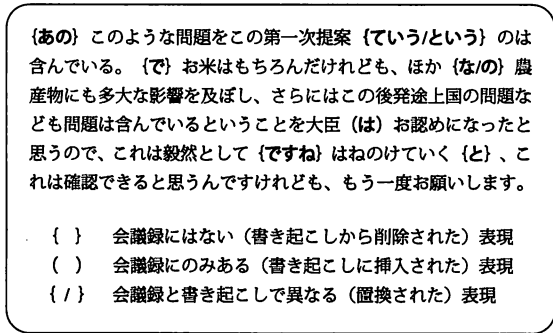


図2 国会音声のタグ付き書き起こしの例

Fig. 2 An example of tagged transcript of meeting

ズや形状、音響設備が異なるため、音響特性も異なったものとなっている。本コーパスではこれらもある程度カバーできるよう会議を選択している。現時点でコーパスは2003年から2007年間の61会議で構成されている。総時間は236時間であるが、これには質問と答弁の間の無発話時間も含まれており、実際にモデルの学習に利用できる時間はやや少ない。

衆議院の会議では会議録が公開されるので、これを音声と照らし合わせながら書き起こすことで、会議録との相違点がタグ付けされた書き起こしを作成した。書き起こしの例を図2に示す。このデータは通常音響モデル・言語モデルの学習に利用でき、また実際の発話内容と会議録との違いを分析することも可能である。このほか、衆議院は第145回国会(1999年)以降の会議録を機械可読な形式で一般に公開しており、これもシステムの開発に利用している。なお、衆議院では会議録の表記に詳細な規則を設けて統一しており、表記の揺れはみられない。

3.2 音響モデル

本システムでは、音響特徴量としてメル周波数ケプストラム(MFCC)とその1次差分(Δ)・2次差分($\Delta\Delta$)各12次元、およびエネルギーの1次・2次差分の計38次元を利用する。特徴量にはケプストラム平均正規化(CMN, Cepstrum Mean Normalization)とケプストラム分散正規化(CVN, Cepstrum Variance

Normalization)を適用している。また話者による音響特徴の偏りをさらに補正するため、声道長正規化(VTLN, Vocal Tract Length Normalization)も行っている。なお、これらの正規化処理の単位は同一話者の発話が連続する区間(話者区間)である。

音響モデルは3状態・Left-to-Right型の対角共分散HMMによるtriphoneモデルである。HMMの統計学習では学習データにおける尤度の最大化を基準とする最尤(ML, Maximum Likelihood)推定が一般的に行われているが、尤度最大化は必ずしも識別能力の最大化とはならないことが問題として指摘されている。これに対して、本システムでは識別学習の1つである最小音素誤り(MPE, Minimum Phone Error)学習[17]を導入した[18]。MPE学習では目的関数 F は式(1)により表される。

$$F(\lambda) = \sum_r \frac{\sum_s p_\lambda(O_r|s) \kappa P(s) \text{RawAcc}(p)}{\sum_s p_\lambda(O_r|s) \kappa P(s)} \quad (1)$$

O_r は観測(学習)データ系列、 $P(s)$ は文候補の単語列 s に対する言語スコア、 κ はスケール係数である。 $\text{RawAcc}(p)$ は文候補における音素正解精度の見積もりであるが、学習時にすべての候補における正確な音素正解精度を算出することは困難であるため、正解音素との時間的なオーバーラップの度合いにより近似的に計算される。ここで分母や分子の和は本来出現しうる全ての文候補をもとに計算されるべきであるが、これは非現実的である。したがって本研究では音声をいったんデコードすることで多数の文候補(正解に対する競合候補)を生成してこれを近似する。なお、これらの文候補はラティスとして保持されている。MPE学習では、正解ラティスおよび競合候補ラティス上でforward-backwardアルゴリズムを適用し、十分統計量を収集する。

3.3 言語モデル

統計的言語モデルは発話の忠実な書き起こしをできるだけ多く収集して構築することが望ましいが、話し言葉音声認識のタスクでは現実的ではない。一方、会議録のような文書は比較的容易に大量入手が可能である。そこで我々は、少量の書き起こしとその正書体が対となったパラレルコーパスから話し言葉スタイルへの統計的変換モデルを学習して、会議録のような文書調のテキストから話し言葉の言語モデルを構築する手法を提案している[19]。この手法では、話し言葉の表現に関してパラレルコーパス中の書き起こしと正書体で相違点を検出し、その統計頻度をもとに確率的な変換パターンを定めて変換モデルを構成する。この変換モデルを用いて正書体の言語モデルから話し言葉スタイルのモデルを生成する。本稿の場合、パラレルコーパスとして利用するのは衆議院審議コーパスにおける書き起こしと会議録である。

言語モデルの変換では、正書体テキスト(W)と話し言葉テキスト(V)をそれぞれ別の言語としてとらえる。このとき、ベイズ則によりそれぞれの言語モデル $P(W) \cdot P(V)$ について次の関係が成立する。

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

ここで $P(V|W)/P(W|V)$ が変換モデルに相当する。したがって、変換モデルの学習では、 V と W の表現の対応（変換パターン）、およびその確率をパラレルコーパスを用いて求めることとなる。変換モデルを適用する際は、言語モデルの代わりに N -gram 単語列を入力し、変換パターンをもとに話し言葉の N -gram 単語列を予測するとともにその頻度を変換確率を利用して推定する。これにより得られた N -gram エントリと頻度をもとに話し言葉の言語モデルを構築する。

本手法における変換モデルは言語モデルと比較すると小規模であるため、より少ない話し言葉データで構築可能であるという特色がある。これに加えて、本手法では同一ドメインのコーパスのみで言語モデルを構築することになるため、日本語話し言葉コーパスや衆議院審議コーパスのようにコーパス内で表記の基準が厳密に守られている場合は表記ゆれが発生しないという利点もあり、会議録作成支援の観点からは好都合である。これに対して、多くの話し言葉音声認識で用いられているコーパス混合・補間では表記がコーパス間で統一されていないことが多く、慎重な対処が必要となる。

本研究ではこれまで 3-gram 統計言語モデルを用いてシステムを構築してきたが、重要な話題語に同音異義語の音声認識誤りが発生するなどの問題がみられた。これに対して、言語モデルの学習データが蓄積されてきたこともあり、本稿では 4-gram モデルを検討する。より長い文脈による詳細な確率モデルとなることで、このような認識誤りを抑制することを期待している。

3.4 発音辞書

発音辞書に関して話し言葉の発音を収集することは容易ではないことから、言語モデルと同様に、統計的変換モデルによる発音の予測手法を我々は提案している [20]。

発音辞書の変換では、音素列の変換（書き換え）パターンからなる変換モデルが用いられる。変換パターンの学習では、まず形態素解析器を用いて標準的な読み（baseform）を学習データ（パラレルコーパス）に付与する。次にこの読みと書き起こし中の実際の読みとの間で音素単位のアライメントを行い発音の変動箇所を同定する。これによる変動前と変動後の音素列の組を変換パターンとして抽出し、学習データにおける生起頻度をもとに変換確率を定める。音声認識用の発音辞書に対しては、これらの変動規則を用いて新たな発音エントリ（surface form）を追加するとともに、発音確率を書き換え確率に基づき推定する。なお、本稿では発音辞書の変換モデルについては衆議院審議コーパスではなく日本語話し言葉コーパスを利用している。

4. 評価実験

4.1 実験条件

これまでに述べた国会音声認識システムについて、認識性能の評価を行った。評価に利用したテストセットは衆議院審議コーパスに含まれる 2007 年分の会議音声（8 会議）である。デコーダは Julius-4.1 を利用し、Intel Core2 Extreme Q6850 プロ

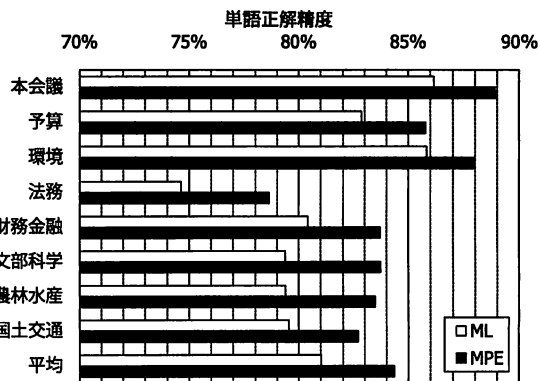


図3 ML 学習と MPE 学習の音響モデルによる各会議の単語正解精度
Fig.3 Word recognition accuracy of test-set meetings by ML-based and MPE-based acoustic models

セッサ（3.0GHz）を搭載した計算機で認識時間が実時間の 3 倍程度になるようにビーム幅等の認識パラメータを設定している。

音響モデルは 2005 年までの審議コーパスの音声データ（134 時間）を利用して学習した。HMM の状態数は 3,000、混合数は 16 である。MPE 学習の効果を明らかにするため、MPE 学習によるモデルと ML 学習によるモデルについて評価を行った。

言語モデルの構築は統計的スタイル変換によって行うが、変換元となる学習テキストは 1999 年～2007 年（テストセット以前）の衆議院会議録である。一方変換モデルの学習には審議コーパスの 2003 年～2006 年分を利用した。テキストの形態素解析は ChaSen-2.4.4+Unidic-1.3.9 [21] により行い、これによる変換元テキストと変換モデル学習テキストの単語数はそれぞれ 161M、2.8M となっている。これらを用いて 3-gram モデルと 4-gram モデルを構築し、性能の評価を行った。なお、言語モデルの語彙サイズは 53,042、これより作成した基本形（baseform）の発音辞書のサイズは 56,348、基本形の辞書から変換モデルにより予測した、音声認識に利用する発音変動形（surface form）を含む辞書のサイズは 62,045 である。

4.2 実験結果

まず、MPE 学習と ML 学習を行った音響モデルによる各会議の単語正解精度を図 3 に示す。このときの言語モデルは 3-gram モデルを利用している。MPE 学習の導入により、単語正解精度が絶対値で 3.3% 向上し、大きな改善を得ることができた。特に法務委員会や文部科学委員会など、テストセットの中でも精度が低めの会議でより大きな改善が得られており、精度の底上げの効果があったといえる。

次に、3-gram 言語モデルと 4-gram 言語モデルによる各会議のパープレキシティ（PP）と未知語率を表 1 に、単語正解精度を図 4 に、文字正解精度を図 5 に示す。ここで利用した音響モデルは MPE 学習を行ったモデルである。表 1 より、言語モデルの文脈長を拡大してもパープレキシティは本会議を除いてほぼ同等となっている。話し言葉表現の比較的小さい本会議で

表1 各会議のパープレキシティと未知語率

Table 1 Perplexity and out-of-vocabulary rates on test-set meetings

委員会	単語数	PP (3-gram)	PP (4-gram)	未知語率
本会議 (10月4日)	24,521	59.00	51.04	0.13%
予算 (10月9日)	50,123	71.97	72.19	0.14%
環境 (10月23日)	31,794	70.16	70.02	0.16%
法務 (10月24日)	44,039	86.45	86.26	0.22%
財務金融 (10月24日)	36,241	68.80	68.93	0.13%
文部科学 (10月24日)	59,307	69.96	69.27	0.14%
農林水産 (10月24日)	37,214	72.04	71.10	0.10%
国土交通 (10月24日)	48,523	73.14	73.74	0.12%
合計	331,762	71.44	70.32	0.14%

は改善がみられていることから、変換モデルにより話し言葉の4-gram が十分に予測できなかった可能性がある。これに対して単語正解精度(図4)と文字正解精度(図5)では、ともに4-gram モデルにより平均で0.4%(絶対値)の改善がみられた。文脈を長くすることで不適切な文仮説が抑制されたと考えられ、4-gram の導入により一定の効果は得られたといえる。MPE 学習による音響モデルと4-gram 言語モデルにより得られた平均の単語正解精度と文字正解精度はそれぞれ84.8%・87.1%となり、本研究の目標とする水準に達しつつある。

なお、本実験では従来の実験[19]より変換モデルの学習データが約4倍に増加しているが、これにより得られた変換パターンの種類はこのときのものと同程度であった。形態素解析システムが異なるため単純に比較はできないものの、これは変換モデルが比較的少量のデータで十分に学習できることを示唆している。

5. おわりに

本稿では、我々が取り組みを進めている国会審議のための音声認識システムについて報告した。国会審議の音声データをもとに、各種の正規化とMPE学習を適用して構築した音響モデル、および統計的スタイル変換に基づいて構築した4-gram 言語モデルと発音辞書から本システムは構成される。実際の会議音声における評価では、85%の単語正解精度、87%の文字正解精度が得られ、性能が実用的な水準に達しつつあることが示された。

謝辞 本研究の一部は、総務省戦略的情報通信研究開発推進制度(SCOPE)「音声認識技術を用いた会議録及び字幕の作成支援システム」により実施された。

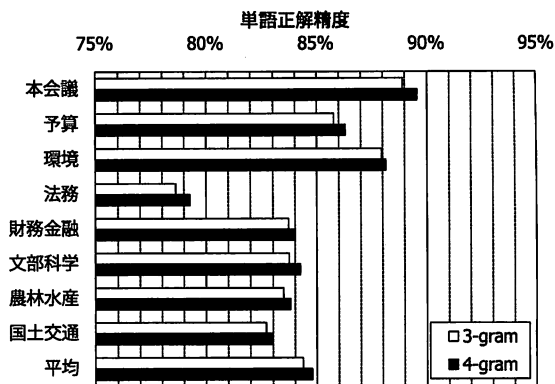


図4 3-gram モデルと4-gram モデルによる各会議の単語正解精度
Fig. 4 Word recognition accuracy of test-set meetings by 3-gram and 4-gram models

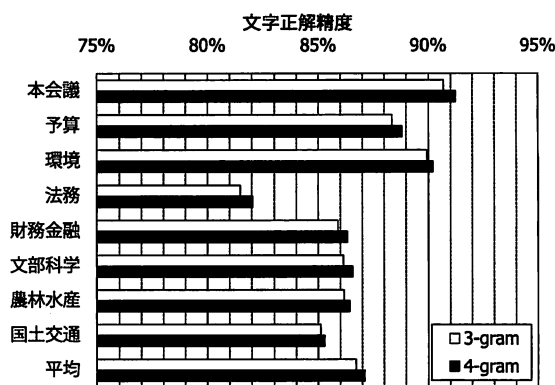


図5 3-gram モデルと4-gram モデルによる各会議の文字正解精度
Fig. 5 Character recognition accuracy of test-set meetings by 3-gram and 4-gram models

文献

- [1] 富樫慎吾, 山口優, 北岡教英, 中川聖一. 講義音声の認識・要約・インデックス化の検討. 情報処理学会研究報告, 2006-SLP-62-11, 2006.
- [2] 根本雄介, 河原達也, 秋田祐哉. スライド情報を用いた言語モデル適応による講義の音声認識と字幕付与. 情報処理学会研究報告, 2007-SLP-66-16, 2007.
- [3] J. Glass, T.J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. Recent Progress in the MIT Spoken Lecture Processing Project. In *Proc. Interspeech*, pp. 2553–2556, 2007.
- [4] S. Furui, K. Maekawa, and H. Isahara. Toward the Realization of Spontaneous Speech Recognition —Introduction of a Japanese Priority Program and Preliminary Results—. In *Proc. ICSLP*, pp. 518–521, 2000.
- [5] H. Nanjo and T. Kawahara. Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition. *IEEE Trans. Speech & Audio Proc.*, Vol. 12, No. 4, pp. 391–400, 2004.
- [6] 中村篤, 大庭隆伸, 渡部晋治, 石塚健太郎, 藤本雅清, 堀貴明, エリックマクダモット, 南泰浩. 音声認識システム SOLON の日本語話し言葉コーパスによる評価 (2006 年版). 情報処理学会研究報告, 2006-SLP-64-44, 2006.
- [7] J.S. Garofolo, C.D. Laprun, and J.G. Fiscus. The Rich Transcription 2004 Spring Meeting Recognition Evaluation. In *Proc. ICASSP*

Meeting Recognition Workshop, 2004.

- [8] S. Renals, T. Hain, and H. Bourlard. Recognition and Understanding of Meetings: the AMI and AMIDA Projects. In *Proc. ASRU*, pp. 238–247, 2007.
- [9] R. Prasad, L. Nguyen, R. Schwartz, and J. Makhoul. Automatic Transcription of Courtroom Speech. In *Proc. ICSLP*, pp. 1745–1748, 2002.
- [10] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu. The LIMSI 2006 TC-STAR EPPS Transcription Systems. In *Proc. ICASSP*, Vol. 4, pp. 997–1000, 2007.
- [11] J. Loof, M. Bisani, Ch. Gollan, G. Heigold, Bjorn Hoffmeister, Ch. Plahl, Ralf Schluter, and Hermann Ney. The 2006 RWTH Parliamentary Speeches Transcription System. In *Proc. ICSLP*, pp. 105–108, 2006.
- [12] B. Ramabhadran, Olivier Siohan, L. Mangu, G. Zweig, M. Westphal, H. Schulz, and A. Soneiro. The IBM 2006 Speech Transcription System for European Parliamentary Speeches. In *Proc. ICSLP*, pp. 1225–1228, 2006.
- [13] 秋田祐哉, 河原達也. 会議音声認識のための発音モデル生成と言語モデル適応. 日本音響学会春季研究発表会講演論文集, 1-5-3, 2005.
- [14] 根本雄介, 秋田祐哉, 河原達也. 会議音声の自動話題分割による単語辞書と言語モデルの適応. 情報処理学会研究報告, 2006-SLP-62-12, 2006.
- [15] 河原達也. [招待講演] 筆記録作成のための話し言葉処理技術. 情報処理学会研究報告, 2006-SLP-64-37, 2006.
- [16] 秋田祐哉, 河原達也. 会議録作成のための話し言葉音声認識結果の自動整形. 日本音響学会秋季研究発表会講演論文集, 3-1-11, 2008.
- [17] D. Povey and P.C. Woodland. Minimum Phone Error and I-smoothing for Improved Discriminative Training. In *Proc. ICASSP*, Vol. 1, pp. 105–108, 2002.
- [18] 三村正人, 河原達也. 話し言葉音声認識タスクにおける音素誤り最小化学習 (MPE) の効果. 日本音響学会秋季研究発表会講演論文集, 3-Q-8, 2007.
- [19] Y. Akita and T. Kawahara. Topic-independent Speaking-style Transformation of Language Model for Spontaneous Speech Recognition. In *Proc. ICASSP*, Vol. 4, pp. 33–36, 2007.
- [20] 秋田祐哉, 河原達也. 話し言葉音声認識のための汎用的な統計的発音変動モデル. 電子情報通信学会論文誌, Vol. J88-DII, No. 9, pp. 1780–1789, 2005.
- [21] 伝康晴, 小木曾智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源: 形態素解析用電子化辞書の開発とその応用. 日本語科学, Vol. 22, pp. 101–122, 2007.