社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64 (36)
2006／12／22

〔招待論文〕 筆記録作成のための話し言葉処理技術

河原 達也†

† 京都大学 学術情報メディアセンター
E-mail: †kawahara@i.kyoto-u.ac.jp

あらまし 話し言葉の音声認識技術の進展に伴い、筆記録 (講演録や会議録) の作成支援が有望なアプリケーションとして考えられるようになってきた。その際には、音声認識だけでなく、言い淀みの整形、口語的表現の修正、文などのセグメンテーションを含めて検討を行う必要がある。本稿では、このような処理を含めて自動書き起こしを行うシステムに関する研究について紹介する。まず、話し言葉音声の書き起こしを対象とした主要プロジェクトを概観し、著者らが提案している高次書き起こしシステムの概要を述べる。そして、文境界の検出と自己修復部 (言い直し) の検出について解説する。最後に、今後の課題について述べる。
キーワード 音声認識, 話し言葉処理, 会議録作成

Intelligent Transcription using Spontaneous Speech Processing

Tatsuya KAWAHARA†

† Kyoto University, School of Informatics
E-mail: †kawahara@i.kyoto-u.ac.jp

Abstract With improvement of the spontaneous speech recognition technology, semi-automatic generation of transcripts or document records of lectures and meetings has become one of its promising applications. For this purpose, we need to take into account post-processing that includes cleaning of verbatim transcripts and segmentation into sentence/paragraph units. This article first gives a brief overview of major research activities on spontaneous speech processing, followed by the proposed statistical framework of an intelligent transcription system. Then, several approaches to sentence unit detection and disfluency detection are described. Finally, future works are discussed.
Key words speech recognition, spontaneous speech processing, rich transcription

1. はじめに

音声によるコミュニケーションは、太古より人間どうしの知識伝達・意見交換の根源的な手段であり、電子的媒体が発達した現代においても、新たな知の創造や重要な意志決定は、主にセミナー・ミーティングや会議などの場で行われていると考えられる。ただし、音声は「揮発的」であるので、内容を記録しておく必要が生じる。近年では、音声メディアのままで大容量の蓄積が可能になったものの、閲覧性や検索の利便性の点から、テキストメディアにして記録するのが一般的である。

音声をテキスト化する (Speech-To-Text) システムの研究は従来から行われているものの、多くはディクテーションシステムのようにヒューマンマシンインタフェースを指向したものである。これに対して、ヒューマンヒューマンコミュニケーションを対象とした場合、音声認識自体が困難になり、音響・発音・言語モデルにおける抜本的な検討が必要となるが、これに加えて、音声認識の枠組み自体の再検討も要する。

話し言葉音声は、考えながら発話されるので、まとまりのない文や言い淀みが多く、100%忠実に書き起こしても、かえって読みづらくなる。実際に、講演録や会議録 (これらを総称して「筆記録」と呼ぶ) を作成する際には、整形作業が必要となる。これには、口語的な表現の文書体への修正も含まれる。そもそも音声波形は一次元信号であり、それを単純に書き起こした (音声認識した) ものは単語の長大な系列にすぎない。すなわち、句読点や改行が全くない文書に相当する。しかも、日本語の話し言葉では、「～ですが」、「～でして」などで文が続くことが多く、境界や前後のつながりが必ずしも明確でない。

これらの問題をまとめると、以下のようになる。

- 言い淀みや無機能語の削除
- 口語的表現の訂正・助詞の補完(注1)
- 文や段落などのセグメンテーション (句読点・改行の挿入)

(注1)：速記者の場合は、誤用や言い誤りの訂正、固有名詞の補完なども行うようである。

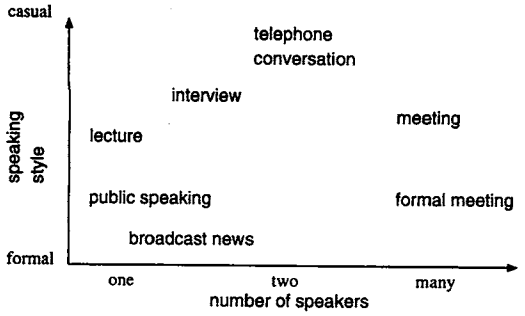

図 1 話し言葉音声認識タスクの分類
Fig. 1 Spontaneous speech recognition tasks

これらの点は、速記者の間でも共通基準とされており、日本速記協会策定の「発言記録作成標準」の整文の基準でも挙げられている。ノートをとったり、議事録を作成する際には、さらに重要文抽出や文圧縮・編集を伴う要約処理を行う必要がある。音声要約に関する研究も盛んになってきた[1]が、本稿では(より忠実な)筆記録を作成することを目標としたアプローチについて概観する。近年、一部の地方議会で音声認識技術を用いた会議録作成支援システムが導入され[2]、裁判所や国会においても導入の動きがある。また講演録においても、最近は、雰囲気を伝える観点から発言に忠実な形式が好まれる傾向もある。ただし、上記の問題を考慮して、システムの設計・研究開発を行う必要があると考えられる。

以下、2章では話し言葉音声を対象とした主要プロジェクトを概観し、3章で著者らが提案する一次整形処理を含めた高次書き起こしシステムの概要を述べる。4章では文境界検出に対するアプローチ、5章では自己修復部(言い直し)の検出について解説する。なお、これらの2つの問題については、DARPA EARS プログラムの Meta-Data Extraction (MDE) タスクでも取り組まれている[3][4][5]。最後に、6章で今後の課題について述べる。なお、話し言葉の音声認識の研究については、[6][7]などを参考にされたい。

2. 話し言葉を対象としたプロジェクトの概観

人間どうしの話し言葉音声を対象とした国内外の主要な研究プロジェクトを、発話スタイルと話者数・話題の広がりの観点から位置づけたものを図1に示す。以下、各々について簡単に説明する。

2.1 放送ニュース

放送ニュースは主要な情報メディアであるので、1990年代半ばから米国の DARPA Hub-4, GALE など多くのプロジェクトで中心的に取り上げられている。また、情報検索や音声要約などの研究においても主な対象となっている[8][9]。大半の発話がプロのアナウンサー/アンカーによるもので、音声認識自体は比較的容易であり、おおむね90%の認識精度となっている。Meta-Data Extraction (MDE) タスクとしても取り組まれているが、言い淀みは非常に少ない反面、文が長いため文境界の検出精度は低い[5]。また日本のNHKでは、ニュースの字幕付与を行うシステムが開発・運用されている[10]。

2.2 講　　演

講演や演説は、記録(講演録)が作成されることが多いので、自動書き起こしに対する需要も大きい。1999年度から5年間にわたる「話し言葉工学」プロジェクトにおいて、660時間に及ぶ学会講演・模擬講演のデータベース『日本語話し言葉コーパス』(CSJ)[11]が構築された。話者はプロの話し手ではないが、公共の場で聴衆に向かって話しているので、発話スタイルは丁寧である。学会講演では、話題や語彙が専門的となる問題があるが、CSJ のテストセットについては約80%の認識精度を達成している。また、CSJ には文境界・言い淀みや重要文・要約などのアノテーションが付与されており、様々な話し言葉処理の研究が行われている[12]。

2.3 講　　義

大学の講義も近年ディジタルアーカイブ化/ネット配信されるようになった。京大の学術情報メディアセンターでは、数百時間にわたって講義の音声・映像アーカイブが蓄積されている。そのため、検索などを指向して音声認識が行われている[13]。また、米国 MIT の数百時間の講義を対象とした iCampus プロジェクト[14]でも、音声認識や検索の研究が行われている。豊橋技科大[15]や東工大でも同様の試みが行われている。ただし、認識精度はいずれも50〜60%にとどまっている。これは、公開の場に比べて、教室で学生相手に話す場合は発話スタイルがよりカジュアルになることと、接話型マイクの使用を教員に強制できないため録音条件が悪いことなどによる。

2.4 一 人 語 り

米国で、ホロコーストの生存者を対象としたインタビューのアーカイブに対して、自動書き起こしを付与する MALACH プロジェクト[16]が行われている。ただし、多くの話者が高齢で、なまりが強く、感情的になることもしばしばあるため、音声認識は容易でない。

2.5 電 話 会 話

米国の DARPA プロジェクト (Hub-5, EARS) では、電話会話を対象とした音声認識が継続的に取り上げられており、Switchboard や Call Home, Fisher などの大規模なデータベースが構築されている。電話会話は密室性・親密度が高いため、上記のいずれに比べても発話スタイルがカジュアルであり、音響的には最も難しい。ただし、専門的な話題や用語は少なく、少数の語彙によるカバレージが大きい。音声認識精度は長らく60%程度であったが、学習データ量の増加(2100時間)や様々な学習法・デコーディング法の組み合わせにより、85%まで改善している[17][18]。Meta-Data Extraction (MDE) タスクとしても取り組まれている。言い淀みは非常に多いが、文境界の検出は比較的容易である[5]。

2.6 ミーティング

近年、ミーティングやグループ討論などもアーカイブ化の対象となりつつある。米国の NIST や欧州の AMI や CHIL などのプロジェクトで、ミーティングを対象としたコーパスの構築と音声認識の研究が行われている[19]。ミーティングでは、多

数の話者が参加するので、誰がいつ発話したかの同定が、インデキシングや音声認識の話者適応において必要となる。これはspeaker diarization タスクと呼ばれるが、ミーティングでは発話の重なりも多く、容易でない。接話型マイクと遠隔マイクの両方で収録を行い、研究・評価を行っており、音声認識精度はおおむね、接話型マイクで 70〜80%、遠隔マイクで 60〜70%程度である。

2.7 公共の会議 (議会)

よりフォーマルで、会議録が作成される会議として、議会がある。欧州では 2004 年度より、EU 議会の音声翻訳を目標とした TC-STAR プロジェクト [20] が行われている。欧州では様々な言語があるので、翻訳の必要性が特に高い。ただし、原稿の読上げが多い本会議を主な対象としており、音声認識精度も約 90%になっている [21]。

著者らは、衆議院の各種委員会を対象として音声認識の研究を行っているが、本会議に比べて自発性が高い予算委員会に対して、マッチドコーパスによる学習を行わない条件で 80%の認識精度となっている。現在、「衆議院審議コーパス」の整備を行い、さらなる向上を図っている。また、我が国のいくつかの地方議会でも音声認識技術の導入が行われており、裁判所での導入に向けた研究開発も進められている。

3. 高次書き起こしシステム

次に、著者らが提案・構築している高次書き起こしシステムについて説明する。

3.1 概　　要

1 章で述べたような一次整形を音声認識の後処理として行う枠組みを図 2 に示す。ここでは、音声認識 (ASR) で得られる忠実な発言体の書き起こし (V) に対して、統計的機械翻訳の枠組みに基づいて、文書体のテキスト (W) に変換するアプローチを採用している [22]。これは、音声認識と同様の定式化になり、ベイズ則により 2 つの確率に分解される。このうち言語モデル $p(W)$ は、文書体の電子化テキストが大規模に存在するので、高い信頼度で推定が可能であり、変換モデル $p(V|W)$ は V と W が対応付けられた比較的少量のコーパス (パラレルコーパス) で学習することになる。

さらに、この枠組みを拡張することにより、音声認識の言語モデル $p(V)$ を推定すること (図 2 下部) を提案している [23]。これは、($p(V)$ 学習用の) 忠実な発言体の書き起こしが非常に少ないのに対して、($p(W)$ 学習用の) 文書体のテキストは大量にあることを利用したものである。ここで、$p(V|W)$ と $p(W|V)$ は、同一の上記パラレルコーパスから学習される。

なお、音声認識と後処理を WFST で実現し、WFST の合成の枠組みで統合的に処理するアプローチも提案されている [24]。統合的な処理による精度・効率の改善が期待されるが、上記の枠組みの方が統計量 (主に言語モデル) の高精度の推定ができると考えられる。

3.2 衆議院審議コーパスによる分析

発言体と文書体の "パラレルコーパス" として、国会討論における実際の発言を忠実に書き起こし、公式の会議録と対応づけた「衆議院審議コーパス」を構築している。本コーパスは、衆議院の予算委員会を中心としながらも、ほぼすべての委員会審議をカバーし、総計 150 時間・200 万単語 (形態素) からなるものである。CSJ の約 1/4 の規模となっている。形態素解析は、茶筌 Ver.2.2.3+IPADIC-2.4.4 に基づいている。

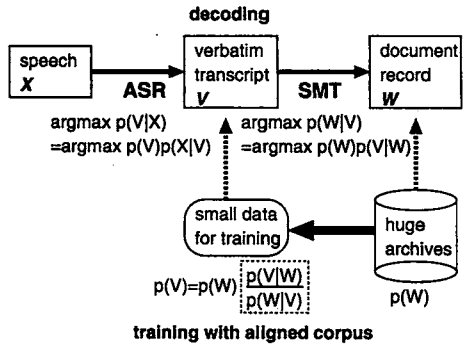

図 2　高次書き起こしシステムの概要
Fig. 2　Overview of proposed transcription system

忠実な書き起こし (発言体:V) と公式の会議録 (文書体:W) の対応づけにおける相違点をモデル化・分析した。前者から後者への変換 ($V \rightarrow W; p(W|V)$)、すなわち整形の過程は、削除・挿入・置換の 3 通りに分類することができる。また、後者から前者への変換 ($W \rightarrow V; p(V|W)$) は、発話の生成過程とみなすことができる。

衆議院審議コーパスにおけるそれぞれの頻度を計数し、高頻度の事例を調べた結果を表 1 に示す。合計で約 11%の形態素で異なりがあることがわかる。ただし、その大半は冗長語の削除である。冗長語としては、いわゆるフィラー (「えー」「あの」など) に加えて、「ですね」「と」などの文末表現もかなり含まれる。言い直しも削除の対象であるが、バリエーションが多いので高頻度にはならない。逆に、会議録において挿入されているのは、格助詞が大半であるが、最も頻度が多かったものは、動詞接尾辞の「い」であった。これは、「してる」「来てる」などの口語的表現の置換ともいえる。(茶筌の形態素単位で) 置換されている口語的表現には、発話の際の息けに起因するもの (「ていう」「けども」など) と、強調的な表現 (「やっぱり」「じゃ」など) があり、前者のパターンの方が多く見られた。

3.3 言語モデルのスタイル変換

これらの現象を統計的にモデル化するには、形態素毎に変換される割合 (確率) を推定するのが直接的な方法であるが、整形の際には、フィラーなどの冗長語は必ず削除され、口語的表現は必ず置換されると考えられるので、これらの変換確率 $p(W|V)$ は 1 とみなし (変換されない場合の W の $p(W) = 0$)、推定する必要はない。表 1 で「要推定」とあるもののみをコーパスから学習する。これらの変換確率は、前後の単語に大きく依存するので、コンテクスト依存のモデルを採用する。しかし、データスパースネスの問題に直面するので、コンテクスト部分に品詞ベースのものも併用することで解決を図っている [23]。

表 1　発言体 (V) と文書体 (W) の主な相違点
Table 1　Major differences between spontaneous speech and document-style text

	$p(W\|V)$	$p(V\|W)$	頻度割合	高頻度上位事例
(フィラー等の) 削除	1	要 推定	8.5%	えー, ですね, あの, まあ, おー, あー, あのー, と, その, で
(助詞等の) 挿入	要 推定	要 推定	1.0%	い {(ex) て (い) る }, は, を, が, と
口語的表現の置換	1	要 推定	1.8%	という/ていう, けれども/けども, いろいろな/いろんな, やはり/やっぱり

衆議院の会議録は膨大な量 (4 年分で 7100 万単語) あるので、その言語モデル $p(W)$ を変換することにより、音声認識用の話し言葉言語モデル $p(V)$ を効果的に構築することができる。これにより、パープレキシティ・認識率の改善効果を得ている。

従来、文書体のテキストと CSJ などの話し言葉コーパスを混合して言語モデルを構築する手法が一般的であるが、このようなコーパスの混合は、対象ドメインに無関係な (CSJ の講演に含まれる) 語彙が多数混入される反面、CSJ に含まれない単語連鎖 (N-gram) には必ずしも話し言葉の特徴が反映されないという問題がある。これに対して提案手法では、対象ドメインにより適合したモデルを構築することができる。

4. 文境界の検出

形態素解析や係り受け解析をはじめとして自然言語処理技術の多くは、文単位の入力を前提としており、また長い講演などの話し言葉に対してインデキシングを行う際にも、文などの単位に区切ることが必要である。本章では、この文境界 (句点) の検出を行う方法について説明する。

4.1 統計的言語モデル

最も基本的な方法は、文境界をポーズの出現した箇所に限定し、その前後のコンテクストを考慮した単語系列 $V = \{w_{-2}, w_{-1}, ポーズ, w_1, w_2\}$ に対して、文境界 (句点) を挿入した場合 $W_1 = \{w_{-2}, w_{-1}, 句点, w_1, w_2\}$ と 挿入しない場合 $W_0 = \{w_{-2}, w_{-1}, w_1, w_2\}$ の尤度 $P(W_1)$ と $P(W_0)$ を比較して、判定するものである。この尤度は、文境界 (句点) が含まれるテキストデータベースから学習した単語 N-gram モデルなどを用いて計算される。この際に、$\log P(W) + \beta(W$ の単語数) といった正規化を行うことが効果的である。ここで、βは単語挿入ペナルティと呼ばれる。これは、音声認識システムにおいて、句点を (ポーズを発音形とするエントリとして) 言語モデルに組み入れて、デコーディングを行う過程に相当する。

ただし話し言葉においては、頻繁にポーズが挿入されるので、単純にすべてのポーズを候補とすると、湧き出し誤りが増加 (適合率が低下) する。また、文境界ほどポーズ長が長くなるといった傾向も観測されず [25]、一方、ポーズがない箇所にも文境界が存在するので、(長い) ポーズのみを文境界候補とすると、検出ミスが増加 (再現率が低下) する。そこで著者らは、前後のコンテクスト (w_{-1}と w_1) に応じて、文境界候補を選別することを提案した [26] [27]。具体的には、文末・文頭表現のある箇所 (ポーズの有無に関わりなく) に文境界候補を限定した上で、話し言葉に特有の「〜と」「〜た」「〜ない」「で〜」の箇所については、ポーズがある場合にのみ対象とする。これは、3.1 節で述べた統計的機械翻訳の枠組みにおいて、変換モデル $p(V|W)$ でポーズの生起を表現／制御することに相当する。

表 2　CSJ における文境界検出結果

	再現率	適合率	F 値
統計的言語モデル	79.2	84.6	81.8
(音声認識結果)	70.2	71.6	70.9
SVM チャンカ	83.0	87.9	85.4
(音声認識結果)	73.9	81.7	77.6

(上段が書き起こし、下段が音声認識結果)

4.2 SVM チャンカ

もう一つの方法として、文境界検出をテキストチャンキングの問題として扱い、連接する単語を同一の文に接続するか、文境界として区切るかの 2 クラス判定を行うサポートベクトルマシン (SVM) を学習することを考える。

SVM チャンカとして YamCha を用い、その入力とする素性として、前後 3 単語の情報 (表記・読み・品詞)、及び話者毎に正規化したポーズ長を利用した。これらは基本的に、前節の統計的言語モデルに基づく手法で用いているものと同じである [27]。

4.3 手法の比較と音声認識結果への適用

上記の 2 つの手法の比較を行った。CSJ において文境界のタグが付与されている 199 講演のうち音声認識のテストセット 30 講演を評価に用いて、残りの講演で学習を行った。また、書き起こしを用いた場合と音声認識を行った場合の比較を行い、認識誤りの影響を調べた。結果を表 2 に示す。

両手法を比較すると、SVM チャンカの方が有意に高い検出性能を実現している。これは、統計的言語モデルと比べて SVM チャンカの方が、直接的に文境界検出を指向して学習が行われているためと考えられる。また、音声認識結果に対する精度は、書き起こしのものに比べていずれも 10%程度低くなっているものの、音声認識誤り率 (約 30%) に比べて低下の度合いは小さい。また、SVM チャンカの方がより頑健であり、80%に近い精度を実現している。これは、N-gram モデルでは 1 つの誤認識が連鎖して他に影響するのに対して、SVM の素性は独立であるためと考えられる。

4.4 係り受け情報の利用

文境界に加えて、係り受けの情報は日本語の構造を規定する上で基本的なものであり、翻訳や要約などのより高度な自然言語処理で利用される。CSJ のコアに対しても係り受けの情報が付与されている。そこで、文境界検出において係り受けの情報を利用したり、係り受け解析と文や節の境界検出を同時並行的に行うことが研究されている [28] [29]。

著者らも、文境界検出と係り受け解析を相互作用的に行う方

式を検討してきた[30]。統計的言語モデルを用いる文境界検出においては、係り受け解析の結果で“区切れている”と思われる(他の文節が自分に係り、自分を飛び越える係り受けがなく、自分が後続に係る確率が低い)箇所を、ポーズの箇所に加えて、文境界候補とした。その結果、適合率が低下したものの、再現率が大きく向上し、F値で2%程度の改善を得た。また、SVMチャンカに基づく手法において、入力の素性として、自分に係る文節の個数などの情報を追加した場合は、F値で0.5%程度の改善であった。

4.5 より多様な単位へのチャンキング

上記のように、文境界の検出精度はおおむね80%前後である。これらの単位が意味的な判断を伴ってトップダウンに設定されたものであるので、自動検出による精度には限界がある。逆に、字幕付与などのアプリケーションを考えた場合には、文より細かい単位に、安定して区切ることが必要である。

そこで、完全にボトムアップに均質な単位にチャンキングを行うことも研究している。具体的には、まず隣接文節間の係り受けの情報を用いて構成要素にチャンキングし、次にこれをポーズまたはフィラーの有無によってチャンキングすることによりフレーズを生成する。前者が節境界を、後者が文境界を、おおむねカバーするもの(再現率93%以上)となっている[31]。

5. 言い淀み・自己修復部の検出

話し言葉音声における顕著な特徴として、フィラーや言い直しなどの言い淀み(disfluency)の現象があるが、従来の自然言語処理技術を適用したり、書き起こしを講演録や会議録の形で保存する際には、これらの部分を削除・修正する必要がある。

「えーと」や「あのー」などのフィラーは、音声認識システムの単語辞書や言語モデルに組込まれているので、フィラーの検出は音声認識の結果として得られる。また、フィラーとして検出されれば、自動的に削除できる(表1で$p(W|V)=1$)。

これに対して、言い直しや繰返しなどの自己修復部の検出は、より複雑な処理を要する。自己修復部に関する最も代表的なモデルは、RIM(Repair Interval Model)[32]である。これは、自己修復部を、被修復部(RPD:ReParanDum)、言い淀み(DF:DisFluency)、修復部(RP:RePair)の3つの区間から構成されると仮定し、言い淀みを検出してから、被修復部を同定し、修復部で置き換えるものである。最近では、自己修復部を“noisy channel model”でモデル化し、3.1節で述べたような、統計的機械翻訳の枠組みで扱うアプローチも提案されている[33][34]。ただし、これらは基本的に、言い淀みが検出され、被修復部と修復部の対応付けがとれることが前提となっている。

しかしながら、CSJのような日本語の長い独話においては、言い淀みが存在しなかったり、被修復部と修復部の対応付けが容易でない場合などの、RIMでは説明できない事例が数多く見られる。そこで、任意の文節に対して、前後の文節のコンテクストから自己修復部(Dタグ)であるかの判定を機械学習により行えるか調べた[35]。ここでも、SVMに基づくYamChaを用い、素性として、直後のポーズの有無、品詞情報、係り先の文節(修復部)と単語が一致する割合などを用いた。ただし、

表3 自己修復部(Dタグ)検出結果

条件		再現率	適合率	F値
部分一致	あり	72.3	82.6	77.1
	なし	13.2	41.2	20.0
係り受け解析	人手	50.7	75.4	60.6
	自動	26.2	54.7	35.4

被修復部と修復部で単語が部分一致するか否かが最も重要な特徴であると予想されるので、学習においても、評価段階においても、(係り先の文節と)部分一致があるか否かでデータを分類した上で行った。

(単語が部分一致している場合)
そういう [風な] 風に 考えられるんじゃないかと
(単語が部分一致していない場合)
[ちょっと穴は] んー 溝は 作れないかもしれない

CSJにおいて係り受け情報・Dタグが付与されている187講演のうち20講演をテストセットとし、残りを学習データとした。さらに、係り受け解析、この場合は被修復部と修復部との対応付けを、CSJのタグ(人手)を用いた場合と自動で行った場合で比較を行った。このような文節間の係り受け解析は難しく、自動解析精度は45.7%しか得られていない。これらの結果を表3に示す。

部分一致している場合はかなり高い精度が得られたが、そうでない場合はほとんど検出ができなかった。また、係り受け解析を自動にした場合は、大きく検出精度が低下した。これらの結果から、自己修復部の検出においては、部分一致の情報がきわめて重要であることがわかる。言い換えれば、表層的な情報のみを用いることの限界を示しているとも考えられる。なお、自己修復部の修正についても検討を行っている[35]。

6. 今後の課題

本稿では、筆記録作成を指向した話し言葉処理について紹介した。音声認識には誤りが不可避であるので、筆記録を作成する際には、人手による修正作業が必要となる。熟練した作業者を想定して、修正を効率的に行うインタフェースの開発[36]とともに、どの程度の音声認識精度を確保する必要があるかの評価・検証が課題となっている。その際には、複数候補や信頼度尺度の利用が有効であるかの検討も行う必要がある。例えば、言い直しなどは信頼度尺度により棄却できる可能性もあり、修正インタフェースの枠組みで検討するのが現実的である。また、修正を行う際には、何らかのセグメンテーションを行う必要があるが、その単位に関する検討も要する。

さらに、講演録や会議録だけでなく、字幕付与やノートテイク支援なども今後の課題に挙げられる。ノートテイクとは、聴覚障害のある学生のために字幕に近いメモをとるもので、大学では近年、その支援が喫緊の課題となっている。これらにおいては、実時間処理が必要で、かつ可読性確保のために圧縮・要約などの処理も必要となるが、完璧な書き起こしでなくても許容されるので、音声認識技術利用の可能性があると考えられる。

謝　辞

本稿で述べた研究成果は、秋田祐哉、高梨克也 (以上、京大)、内元清貴 (NICT) ら各氏の多大な貢献によるものであり、深く感謝します。また、日頃よりご協力頂いている衆議院事務局に感謝します。

文　献

[1] S.Furui. Recent advances in automatic speech summarization. In *Proc. IEEE Workshop Spoken Language Technology*, 2006.

[2] 山崎恵喜. 音声認識システムを活用した会議録作成 –北海道議会における実例–. 情報管理, Vol. 49, No. 4, pp. 165–173, 2006.

[3] Y.Liu, E.Shriberg, A.Stolcke, B.Peskin, J.Ang, D.Hillard, M.Ostendorf, M.Tomalin, P.Woodland, and M.Harper. Structural metadata research in the EARS program. In *Proc. IEEE-ICASSP*, Vol. 5, pp. 957–960, 2005.

[4] E.Shriberg. Spontaneous speech: How people really talk and why engineers should care. In *Proc. INTERSPEECH*, pp. 1781–1784, 2005.

[5] Y.Liu, E.Shriberg, A.Stolcke, D.Hillard, M.Ostendorf, and M.Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Trans. Audio, Speech & Language Process.*, Vol. 14, No. 5, pp. 1526–1540, 2006.

[6] 河原達也. (サーベイ) 話し言葉音声認識の概観. 電子情報通信学会技術研究報告, SP2000-95, NLC2000-47 (SLP-34-21), 2000.

[7] 河原達也. 『日本語話し言葉コーパス』を用いた音声認識の進展. 話し言葉の科学と工学ワークショップ, pp. 61–66, 2004.

[8] B.Chen and L.-S.Lee. Spoken document understanding and organization. *Signal Processing Magazine*, Vol. 22, No. 5, pp. 42–60, 2005.

[9] C.Cieri, D.Graff, M.Liberman, N.Martey, and S.Strassel. The TDT-2 text and speech corpus. In *DARPA Broadcast News Workshop*, pp. 57–60, 1999.

[10] 安藤彰男. ニュース音声自動字幕化システム. 情処学研報, 2000-SLP-34-28, 2000.

[11] 前川喜久雄. 『日本語話し言葉コーパス』の概観. In `http://www2.kokken.go.jp/~csj/public/members_only/manuals/overview10.pdf`, 2004.

[12] 河原達也. (招待講演) CSJ を用いた話し言葉の音声認識・言語解析の進展. 日本音響学会研究発表会講演論文集, 3-1-6, 春季 2006.

[13] 北出祐, 河原達也. 講義の自動アーカイブ化のためのスライドと発話の対応付け. 情報処理学会研究報告, SLP-55-11, 2005.

[14] A.Park, T.Hazen, and J.Glass. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In *Proc. IEEE-ICASSP*, Vol. 1, pp. 497–500, 2005.

[15] 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.

[16] B.Ramabhadran, J.Huang, and M.Picheny. Towards automatic transcription of large spoken archives - English ASR for the MALACH project. In *Proc. IEEE-ICASSP*, Vol. 1, pp. 216–219, 2003.

[17] S.F.Chen, B.Kingsbury, L.Mangu, D.Povey, G.Saon, H.Soltau, and G.Zweig. Advances in speech transcription at IBM under the DARPA EARS program. *IEEE Trans. Audio, Speech & Language Process.*, Vol. 14, No. 5, pp. 1596–1608, 2006.

[18] S.Matsoukas, J.-L.Gauvain, G.Adda, T.Colthurst, Chia-Lin Kao, O.Kimball, L.Lamel, F.Lefevre, J.Z.Ma, J.Makhoul, L.Nguyen, R.Prasad, R.Schwartz, H.Schwenk, and B.Xiang. Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system. *IEEE Trans. Audio, Speech & Language Process.*, Vol. 14, No. 5, pp. 1541–1556, 2006.

[19] J.S.Garofol, C.D.Laprun, and J.G.Fiscus. The RT-04 spring meeting recognition evaluation. In *NIST Meeting Recognition Workshop*, 2004.

[20] 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.

[21] J.Loeoef, M.Bisani, C.Gollan, G.Heigold, B.Hoffmeister, C.Plahl, R.Schlueter, and H.Ney. The 2006 RWTH parliamentary speeches transcription system. In *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, pp. 133–138, 2006.

[22] 下岡和也, 南條浩輝, 河原達也. 講演の書き起こしに対する統計的手法を用いた文体の整形. 自然言語処理, Vol. 11, No. 2, pp. 67–83, 2004.

[23] 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.

[24] T.Hori, D.Willett, and Y.Minami. Language model adaptation using WFST-based speaking-style translation. In *Proc. IEEE-ICASSP*, Vol. 1, pp. 228–231, 2003.

[25] 柏岡秀紀. 独話データのポーズ単位を利用した節境界判定. 情報処理学会研究報告, 2005-SLP-57-15, 2005.

[26] 下岡和也, 秋田祐哉, 南條浩輝, 西光雅弘, 河原達也. CSJ の文境界推定における統計的言語モデルと SVM の比較・評価. 日本音響学会研究発表会講演論文集, 3-1-2, 春季 2006.

[27] Y.Akita, M.Saikou, H.Nanjo, and T.Kawahara. Sentence boundary detection of spontaneous Japanese using statistical language model and support vector machines. In *Proc. INTERSPEECH*, pp. 1033–1036, 2006.

[28] 大野誠寛, 松原茂樹, 柏岡秀紀, 加藤直人, 稲垣康善. 節境界単位での漸進的な独話係り受け解析. 情報処理学会研究報告, 2005-SLP-55-10, 2005.

[29] 大庭隆伸, 堀貴明, 中村篤. 自然発話理解のための逐次的係り受け解析. 音講論, 3-7-10, 秋季 2005.

[30] 下岡和也, 内元清貴, 河原達也, 井佐原均. 日本語話し言葉の係り受け解析と文境界推定の相互作用による高精度化. 自然言語処理, Vol. 12, No. 3, pp. 3–17, 2005.

[31] 西光雅弘, 河原達也, 高梨克也. 隣接文節間の係り受け情報に着目した話し言葉のチャンキングの評価. 情報処理学会研究報告, SLP-61-4, 2006.

[32] C.Nakatani and J.Hirschberg. A speech first model for repair detectin and correction. In *Proc. ARPA Human Language Technology Workshop*, pp. 329–334, 1993.

[33] M.Honal and T.Schultz. Correction of disfluencies in spontaneous speech using as noisy-channel approach. In *Proc. EUROSPEECH*, pp. 2781–2784, 2003.

[34] S.Maskey, B.Zhou, and Y.Gao. A phrase-level machine translation approach for disfluency detection using weighted finite state transducers. In *Proc. INTERSPEECH*, pp. 749–752, 2006.

[35] 下岡和也, 河原達也, 内元清貴, 井佐原均. 『日本語話し言葉コーパス』における自己修復部 (D タグ) の自動検出および修正に関する検討. 情報処理学会研究報告, SLP-56-14, NL-167-14, 2005.

[36] 南條浩輝, 秋田祐哉, 河原達也. 音声認識を利用した会議録・講演録の作成支援システムの設計と評価. 日本音響学会研究発表会講演論文集, 1-7-13, 秋季 2005.