メディア教育研究　第９巻　第１号
Journal of Multimedia Education Research 2012, Vol.9, No.1, S1－S8

特集（招待論文）

話し言葉の音声認識の進展
―議会の会議録作成から講演・講義の字幕付与へ―

河原　達也[1)]

音声認識技術は，この十年余りの間に大きな進歩を遂げている。講演・講義や議会審議などの公共の場で話される音声に対しても研究開発が進められ，一部は実用的なレベルに達しつつある。本稿では，このような話し言葉を対象とした音声認識の最近の技術動向と応用対象を解説するとともに，講演・講義への字幕付与やノートテイク支援など，メディア教育への展開について紹介を行う。

キーワード
音声認識，話し言葉，字幕付与，ノートテイク，会議録作成

1．はじめに

音声認識の研究が開始されたのは今から50年以上も前に遡る。京都大学では1960年頃に「音声タイプ」が作成された（Sakai & Doshita, 1962）。これは，今の大型計算機ほどの大きさの真空管／トランジスタの回路で，単音節（「あ・お・い」など）の認識を行うものであった。その後30年くらいは，音声認識に有効な音響特徴量と動的パターンのマッチング手法に関する基礎的な研究が世界中で行われた。そして，現代の音声認識システムの原型ができたのは1990年頃である。これは，スペクトル包絡を表現する特徴量（ケプストラム）と統計的分布の状態遷移モデル（HMM: Hidden Markov Model；隠れマルコフモデル）に基づくものである（古井，2009）。それ以降約20年が経過したが，音声認識の基本的な枠組みは変わっていない。

しかし，音声認識技術はその間に飛躍的に進歩した。これは，モデルの洗練とデータの大規模化によるものである。その間の計算機の処理能力の大きな向上によるところもある。1990年代から2000年代半ばにかけては，パソコンや携帯端末の高性能化に伴って，それらの端末機器で動作するように設計・開発されてきたが，その後ネットワークが高速になるとサーバ・クラウド型のシステムが実現された。その結果，従来は考えられなかったような超大規模なデータに基づく高精度な処理が実現された。音声認識は10年ほど前からカーナビやパソコンソフト等で実用化されているが，現在スマートフォン等で動作しているものは格段に性能が高くなっている。

ただしこれらはあくまで，人間が機械に向かって話すヒューマン・マシン・インタフェースとしての位置づけである。すなわち，ユーザは話す内容を事前に考えて（概念的制約），文法的で単純な文を（言語的制約），明瞭に発声する（音響的制約）必要がある。これは例えて言うと，我々が外国に旅行に行ってホテルやレストランで要求や情報提示などのコミュニケーションを行っている話し方に近い。これに対して，母国語の人間どうしが行っているコミュニケーションでは，そのような単純なものだけでなく，様々な知識の伝達や深い議論を行っている。この場合，考えながら発話を行うため，言語的にも音響的にも明瞭とは限らない（図１参照）。このような人間どうしの話し言葉を対象とした音声の認識についても研究が行われ，対象範囲はまだ限られているが徐々に実用的なレベルになっている。本稿では，この研究開発に関する最近の動向とともに，メディア教育への展開について紹介を行う。

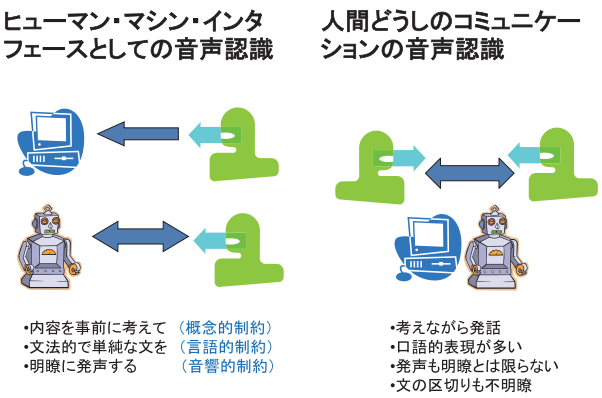

図１　機械相手から人間どうしの音声の認識へ

[1)] 京都大学

2. 音声認識の話し言葉への展開

音声認識の人間どうしの話し言葉コミュニケーションへの展開について図２にまとめる。この図に挙げているのは，これまで音声認識の研究プロジェクトで取り組まれた応用対象である。この図の縦軸は，発話スタイルの丁寧（フォーマル）さを表す。「読み上げ」というのは，与えられた文または事前に考えた文を読み上げている状況で，最も丁寧な発声になる。プロのアナウンサによる放送ニュースも大半は原稿の読み上げである。講演会や議会においては，原稿を読み上げている場合もあるが，大半はそうではない。しかし，公共の場でのスピーチであるので，話す内容の大筋は事前に準備しているし，発声も基本的に明瞭になるように心掛けている。ただし発話が長いので，１文１文丁寧に発声するという感じにはならない。大学の講義は閉じた場であるので，もっとくだけた感じになる。さらに通常の会議やミーティングも同様であるが，話者が複数人になる。図の横軸は主な話者の数である。図の一番上の電話会話やインタビューは，特定の話題について自由に話してもらっているもので，最もくだけたスタイルになる。

上記のうち主要なもの，特に実用的な対象について以下に簡単に述べる。

(1) テレビニュース番組への字幕付与

放送ニュースを対象とした音声認識の研究は，1990年代半ばから米国DARPAプロジェクトで行われており，アナウンサの読み上げ部分については90%程度の認識率が得られている。しかし，生放送の字幕付与に供するには，95%以上の精度が必要とされた。NHK放送技術研究所は，当該アナウンサや当日のニュース原稿に音声認識のモデルを適応することで，世界に先駆けて2000年３月にこのシステムを実現した。アナウンサ以外は復唱入力を導入するなど，その後も改良が重ねられているが，詳細は本特集の記事を参照されたい（佐藤，2012）。

(2) 議会の会議録作成支援

議会審議を対象とした音声認識の研究は，2000年代前半から欧州議会を対象としたTC-STARプロジェクト（Ramabhadran, Siohan & Sethy, 2007）や，日本の国会を対象に京都大学（著者の研究室）で行われてきた。それと並行して2005年頃から，いくつかの地方議会で商用の音声認識ソフトを用いて会議録作成を行うシステムが導入されている。北海道議会の報告によると，原稿の読み上げが多い本会議では80%～90%の精度だが，自由闊達な討論が行われる委員会審議では70%程度であった（山崎，2006）。これに対して著者らは，話し言葉の精緻なモデル化を行い，大規模な審議データに適用することで，国会の委員会審議でも90%に近い認識精度を実現した。

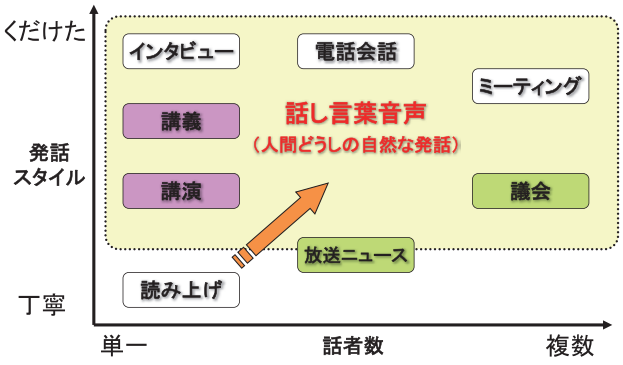

図２　話し言葉の音声認識の応用対象

このシステムは，2010年に衆議院に導入され，2011年から実運用されており（河原，2012；Kawahara，2012；猿谷，2012），国会レベルでは世界初のものである。このシステムの詳細については４章で述べる。

(3) 裁判所の公判の検索と記録

議会審議に類似したものとして裁判所の公判がある。米国では一部の速記者が，商用の音声認識（ディクテーション）ソフトを用いて復唱入力を行っている。裁判は議会と比べて，一般人が発言し，個別的な固有名詞が多いので，音声認識は容易でないと考えられる。我が国では2009年に裁判員制度が導入されたのに伴い，公判が連日開催されるようになり，音声・映像の記録・可視化も導入された。その音声・映像を効率的に検索するために音声認識技術が導入された。このシステムはNECが開発したものである（越仲・江森・大西・北出・谷・佐藤，2010）。その後，公判記録の作成支援にも利用されている。

(4) 講演・講義の書き起こし作成・字幕付与

学術講演を対象とした音声認識の研究としては，1990年代後半から我が国で行われた『日本語話し言葉コーパス』（CSJ）（前川，2004）を用いたものが挙げられる（河原，2006）。その後世界各地で，大学の講義などを含めて研究が行われた。代表的なものにMITのOCW（OpenCourseWare）を対象としたものがある（Glass, Hazen, Cyphers, Malioutov, Huynh & Barzilayet, 2007）。また，最近ではTED（http://www.ted.com/）の講演を対象として音声認識・翻訳を行う試みが行われている（Paul, Federico & Stucker, 2010）。比較的フォーマルな講演では80%程度の認識率が得られているが，大学の講義だと60%～70%程度である。多くのプロトタイプシステムが作成され，一部実証実験なども行われているが，完全に実用レベルに到達したものは現時点でない。詳細は５章で述べる。

表１　話し言葉の音声認識のレベルと応用

音声認識精度	主観レベル	応用
60%～80%	話題・キーワードが把握できる	音声・映像の検索（裁判・講演・講義）
75%～90%	発言内容が把握できる	講演・講義への字幕付与
85%～95%	誤りが散見される	議会の会議録作成
95%～	ほぼ完璧	放送などの字幕付与

上記をふまえて，話し言葉の音声認識のレベルと応用についてまとめたものが表１になる。数値の範囲には若干のマージンがあるが，著者らの経験・知見では，人間が見て書き起こしとして意味をなす最低ラインは認識率75%程度で，それ以下では内容の把握が困難であるばかりか，不快な印象を与える（南條・秋田・河原，2005）。認識率が90%を上回ると，誤りがあってもほとんど理解に支障がなく，「間違い探し」のレベルになる。生放送の字幕付与では認識率95%以上が要件とされているが，原稿の読み上げかプロのアナウンサでないと実現困難である。そもそも自然な話し言葉には，言い淀みや冗長語が５%程度は含まれるので「正解」の定義自体が自明でない。すなわち，言い淀み等を含めて書き起こしても結局修正を行う必要があり，自動で修正を行えないとすると５%程度の誤り率は不可避となる。議会の会議録作成の場合，90%以上の認識率が望まれるが，最低でも80%は必要である。認識率にはばらつきがあるので，大半の音声区間に対して80%を確保しようとすると，平均で85%程度が目安となる。衆議院の場合はこのようにして性能要件が設定された（秋田・三村・河原，2010）。

もちろんある程度認識率が高くなくても，一から入力するよりは，音声認識結果を修正する方が効率よく書き起こしを作成できるが，その場合でも認識率75%程度が最低ラインであろう。認識率がそれより低いと，個々の発言内容を把握するのは困難になる。しかし，全体としてどういう話題を話しているのかは推察できるので，キーワードを元に検索するなどの応用には供することができる。

図２と表１を比較すると，図２の下の方ほど高い認識率が得られることがわかる。議会と講演では，講演の方が容易であるようにも思われるが，議会の方が高い認識率が得られているのは，後述するように音声認識システムのモデルを学習するためのデータが大規模にあるためである。

3.　音声認識の原理と課題

はじめに述べたように，現在の音声認識の基本的な枠組みは1990年頃に確立され，その後世界中のほとんどすべてのシステムで普遍的なものになっている。その一方で，音声認識システムは前章で述べた応用対象毎に，かなりの労力をかけて作成されており，しかも年々進化し続けている。これは例えていうと，自動車に関して，ガソリンエンジン，ギア・シャフト，車体などから構成される枠組みは約百年間変わっていないのに，メーカ毎にしかも客層毎に毎年様々な車種が作られているようなものである（もっとも最近は，電気自動車やハイブリッドカーも出現しているが）。

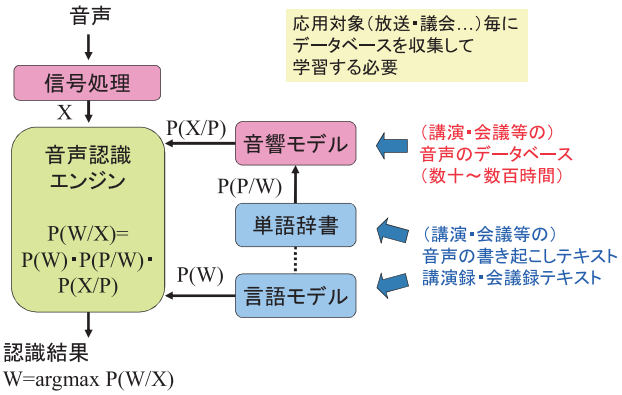

図３　音声認識の原理

3.1　音声認識の原理

音声認識の原理を図３に示す。自動車で性能上最も重要なのがエンジンであるのと同様に，音声認識システムにもエンジンがあり，これが最も高度なプログラムである。音声認識エンジンは必要な認識精度と処理速度を実現する上で重要であり，現在では技術的に高度になりすぎたので，自力で開発できるところは世界中でもそれほど多くない。著者らは1990年代後半から，誰でもどのような目的でも使えるオープンソースの音声認識エンジンJulius（http://julius.sourceforge.jp/）の開発を進めており，国内外で幅広く使用されている。

ただし，応用対象に必要な仕様・性能を実現するのは，主としてエンジンではなく，図３の右側の３つのモジュール（モデル）である。エンジンがソフトウェアプログラムであるのに対して，これらのモデルは巨大なデータベースである。音響モデルは，音素毎の音響特徴量（ケプストラムなど）の分布を記憶する統計モデル（HMM）で，応用対象における音響環境・話者層・発話スタイルに合致するように構築する。すなわち，議会向けのシステムであれば議会審議の音声を，講演を対象としたシステムであれば講演音声を，大規模に収録したデータベースを構築した上で統計量を学習する。単語辞書は，応用対象で出現が想定される単語とその読み（音素表記）のリストである。言語モデルは，それらの単語の連鎖の統計量（N-gramモデル），すなわち単語列の相対頻度を記憶している。これらは，応用対象の話題や発話スタイルに合致するように，議会の会議録や講演の書き起こしな

どのデータベースを構築して学習する。

要するに，話し言葉全般に適用できる音声認識システムが世の中に存在するわけでなく，応用対象毎に合致したモデルを構築する必要があり，このモデルの善し悪しが認識性能を左右する。モデルの善し悪しは，その学習方法にもよるが，最先端の技術を用いたとすると，学習データベースの規模が最も重要になる。

3.2　話し言葉音声認識の課題

上記から，話し言葉の音声認識のための最大の課題は，話し言葉の音声と書き起こしのデータベースの構築に帰着される。データを収集すればよいだけのように思われるが，大規模に行うのは容易ではない。人間が書いた文章は，新聞・論文やWeb上のテキストなど大規模に存在する。しかし，これらは基本的に話し言葉ではない。一方，話し言葉の音声は，日々の講義や会議などを収録すればよい。しかし，これらには通常書き起こしがない。音響モデルを学習するには，数百～数千時間の忠実な書き起こしが付与された音声データが必要である。これだけの分量の音声に対して，言い淀みも含めて忠実に書き起こしを作成するには，膨大な手間とコストが必要となる。さらに，言語モデルを学習するにはこれでも不十分で，書き起こしでなくても，応用対象に沿ったテキストをさらに集める必要があるが，これも容易でない。

このような研究目的のために，『日本語話し言葉コーパス』（CSJ）が構築された。これは，学会講演や模擬講演を計600時間収集し，忠実に書き起こし，言語的なアノテーションを行ったものである。講演の音響モデルを構築するには十分な規模であるが，言語モデル構築用のデータとしてはカバーしている範囲及び規模ともに十分でない。また，会議など他の応用対象に適用するにはミスマッチが大きい。

これに対して，個別の講演・講義・会議などに対して，関連するデータを収集して，モデルを適応するアプローチも考えられる。例えば講義であれば，使用する教科書や講義スライド，さらには同じ講師が過去に行った講義音声などを利用できれば，効果的である。著者のグループでいくつかの講義を対象に評価を行ったところ，CSJのみで学習したモデルでは単語認識率が61%であったが，１回（90分）の講義で言語モデル・音響モデルの適応を行い，さらにスライドから語彙を追加することで10%近く認識率が向上した。平均で約70%であるが，講師によって60%から80%くらいまで異なる。米国・MITのグループでも同様の報告がされている（Glass et al., 2007）。

4.　国会審議の会議録作成支援のための音声認識

議会では長らく手書き速記によって逐語的な会議録が作成されてきたが，今世紀になって速記者の新規養成が停止され，代替手段が模索されてきた。衆議院では，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に対して音声認識を行い，会議録の草稿を生成する。この音声認識の主要モジュール（音響モデル・言語モデル等）に著者らの研究成果が導入されている（河原，2012；Kawahara, 2012）。

4.1　会議録から話し言葉への統計的自動変換

本システムの研究開発に際して鍵となったのは，前章で述べたように，大規模な学習データベースである。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。

しかし，会議録の文章は，実際の発言内容と比べると，「えー」「あのー」などのフィラーや「～ですね」などの冗長な文末表現が削除され，「それじゃ」「～してる」などの口語的表現が「それでは」「～している」に修正されるなど，かなりの差異があり，そのままでは音声認識のモデル学習（正解テキスト）に使用することはできない。従来の枠組みでは，フィラーなどを含めた忠実な書き起こしを手作業で作成する必要があったが，膨大な手間とコストを要し，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。

そこで著者らは，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を考案した（河原，2012；Kawahara，2012）。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４に示す。発言の忠実な書き起こしと会議録を対応づけて分析した結果，13%の単語で違いがみられたが，その93%はフィラーの削除や語句の修正のような単純な編集であった。これらに関しては，統計的な機械翻訳の枠組みでモデル化できる。これにより，10年以上分の会議録（約２億単語）のテキストから，審議で発言される単語系列を予測し，その頻度を推定することで，話し言葉の言語モデルを構築することができた（Akita & Kawahara，2010）（図４の右半分）。

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を行う手法を考案した（図４の左半分）。会議のターン（=発言者が交代するまでの発言区間）毎に，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，これを用いて音声と照合することで，実際に発言された内容の書き起こし（音素ラベル）を復元する。本手法は，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（三村・秋田・

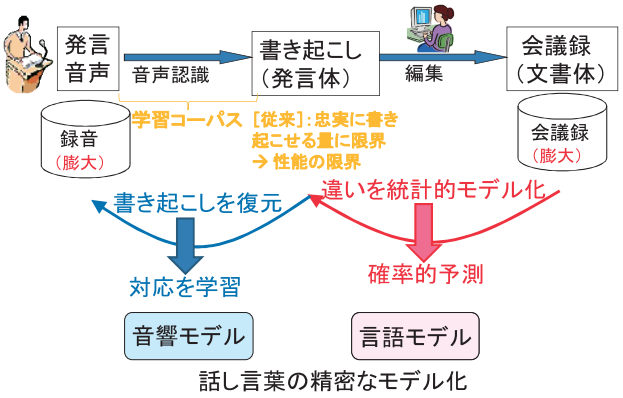

図４　言語モデル変換に基づくモデル学習の枠組み

河原, 2011), 千時間規模の音声データに適用されている。

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更新を可能にするもので, 今後さらに多くのデータが蓄積されることによって，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。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。

4.2　システムの構成と評価

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）に基づく音声認識エンジン（NTTで開発）に統合することで構成された（河原，2012；Kawahara，2012)。

システムは，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入力する。質問者と答弁者（＋議長）には別のチャネルが割り当てられており，システムが自動的にチャネル選択と話者区分化を行い，音声認識を実行する。会議録作成が目的であるので，音声認識は厳密にリアルタイムである必要はないが，会議の最中から会議録作成作業を行うので，音声認識処理の実時間比速度はほぼ１である必要がある。

音声認識結果は，作業単位（通常５分）ごとに原稿作成者（速記者）に割り当てられ，専用のエディタを用いて修正・編集される。原稿作成者が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。エディタは，元音声と映像に時刻・ターン（発言区間）・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。音声認識結果に，自動的にフィラーをマークしたり削除することもできる。ただし，その他の編集の自動化は難しい。

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年度に試行を行った結果，文字正解率は平均89.3%であった。この音声認識結果を，速記者が専用エディタで修正・編集することにより会議録原稿を作成するシステムの有用性が検証され，2011年４月から本格的なシステム運用となった。2011年に行われた118会議で評価したところ，平均文字正解率は89.8%であった。85％を下回る会議はほとんどなく，本会議に限ればほぼ95％に達していた。

ただし, 認識誤りが10％程度存在するのも事実であり，これ以外に言い淀みや口語表現で編集が必要な箇所も10％程度ある。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。

音声認識のモデルは導入後も随時更新している。単語辞書・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，一時的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の際に更新されることになっている。

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が，会議録と審議音声が大規模に集積されているという特性によるところが大きい。

5.　講演・講義への字幕付与

講演や講義などの教育の現場において，情報通信技術（ICT）の導入が進められている。特に，講演・講義の映像・音声を収録して，配信するサービスが徐々に導入されている。著者らは，音声認識技術の適用に関して研究開発を進めている。

講演・講義に字幕付与を行う形態・目的を図５に分類する。形態としては，録画した映像（アーカイブ）に対して後で付与する場合と，講演・講義の最中にリアルタイムに付与する場合に分類される。目的としては，e-Learning用の映像アーカイブに対して，そこで話されている内容に即したインデクスを付与し，検索や効率的視聴に供することと，聴覚障害者・外国人・高齢者等に対して情報保障を行うことが挙げられる。一般の人でも，字幕があることで理解が深まることが期待される。

字幕は誤りがない完璧なものが望まれるが，人手で付与するのは多大な手間とコストを要する。音声認識技術を用いることで省力化・迅速化が期待できる。しかし，音声認識には誤りが不可避で，特に講演や講義のような話し言葉を高い精度で音声認識するのは容易でないので，人手で修正する必要がある。さらに，話し言葉をそのままテキストにしても可読性がよいとは限らず，一定の整形作業も要する。これは会議録作成の場合と同様である。ただし，検索用インデクスのみに用い，字幕そのものを提示しないのであれば，キーワードが認識される

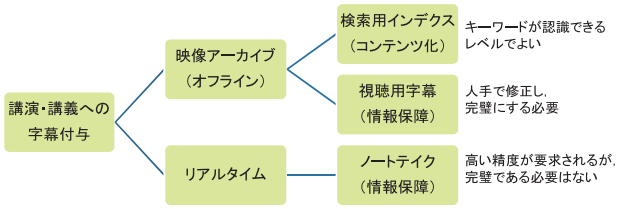

図５　講演・講義への字幕付与の形態と目的

レベルでよいと考えられる（表１参照）。

5.1　講演・講義映像へのオフライン字幕付与

近年，数多くの講演・講義が収録され，映像配信されるようになっている。これには，受講生へのサービスの拡充の観点と，一般社会への広報・発信の観点があり，後者の代表例として，OCW（OpenCourseWare）やiTunesUが挙げられる。特に前者の目的では，単に講義の映像をストリーム配信するだけでなく，使用されたスライドを同期して表示したり，スライドに沿って音声・映像を再生・スキップする機能も提供されることが多い。ただしそのためには，講演・講義を通してスライドを専用のソフトウエアを介して使用してもらう必要がある。

そのような前提が満たされず，スライドに基づいてブラウズ・検索できない場合には，音声認識に基づいて検索のためのインデクスを作成することが検討される。このような検索用途であれば，50%～70%程度の認識率でも十分である（100%正しい書き起こしと同様の検索結果が得られる）ことが示されている（Van Thong, Moreno, Logan, Fidler, Maffey & Moores, 2002）。

一方，字幕として提示する場合には，音声認識結果を修正し，句読点や改行を挿入して，テキストとして整える必要がある。広島大学・アクセシビリティセンターでは，字幕を付与した教材を配信する試みを行っている（山本，2011）。また，このようなユニバーサルな教育支援のための技術開発・実証実験を行う国際的な枠組みとして，Liberated Learning Consortium（http://liberatedlearning.com/）があり，専用のエディタなどを開発している。

米国・MITや日本のいくつかの大学（京都大・豊橋技術科学大・東京工業大など）では，講義の音声認識の研究，及びそれに基づくブラウザの試作を行っている。また，近年TEDの講演を対象に音声認識そして機械翻訳を行う研究プロジェクトも行われている（Paul et al., 2010）。翻訳も実現できれば，外国語の講演も理解しやすくなるので，多くの利用が見込まれる。

ただし前記の通り，音声認識精度は60%～80%であるので，字幕として用いるにはかなりの修正を必要とする。広島大学の調査では，１回90分の講義に対して，認識率

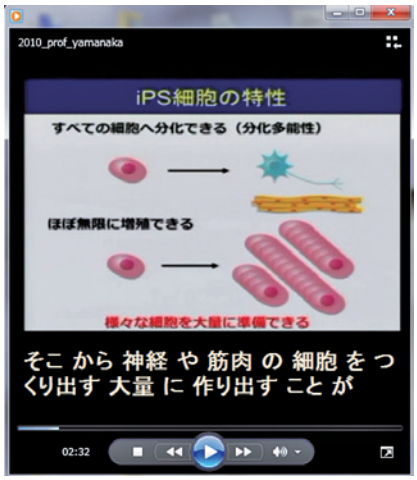

図６　音声認識による講演への字幕付与の例

が70%では10.8時間の編集作業が必要となるが，90%では3.6時間になると報告している（山本，2011）。したがって，90%程度の認識率を確保するために，復唱入力方式を採用している。また，同調査では，字幕の意義は認められた半面，「話し言葉そのままなので読みにくい」などのコメントがあった。

著者らも，京都大学OCWで公開されている講演映像を対象として字幕付与を進めているが，音声認識誤りを修正するだけでなく，話し言葉を読みやすく整形したり，必要最小限の句読点を入れたりする必要性を感じている。これらは，議会の会議録作成の場合と共通の課題であるが，国会の場合はプロの速記者が担当しているのに対し，大学で学生アルバイトなどに作業してもらうことを想定すると，その基準策定や訓練が検討課題である。音声認識により字幕付与した例を図６に示す。これを編集した字幕が，京都大学OCWの講演映像の一部に用いられている。

5.2　教育現場におけるリアルタイム字幕付与

聴覚障害のある学生がいる講義では，その場で情報保障を提供する必要がある。そのため多くの大学で，ボランティア学生によるノートテイクが行われている。その大半は講師の発話内容を紙に書いていく形態であるが，書く速度は話す速度に比べて圧倒的に遅いので，しばしば「２割要約」などと言われる。これに対して，パソコンを用いたノートテイク（ＰＣテイク）も採用されるようになっている。より高速な入力が可能で，２名で連係入力するソフトIPtalk（http://www.geocities.jp/shigeaki_kurita/）を用いると，ほぼすべての発話内容を字幕化することができる（吉川・太田・白澤，2001）（図７(a)）。ただし，いずれの場合も長時間作業できないので，５分～10分毎に交代しながら行うのが一般的である。したがって，かなり大がかりな人数及び装置が必要になる。

さらに最近は，音声認識を用いた方式も模索されてい

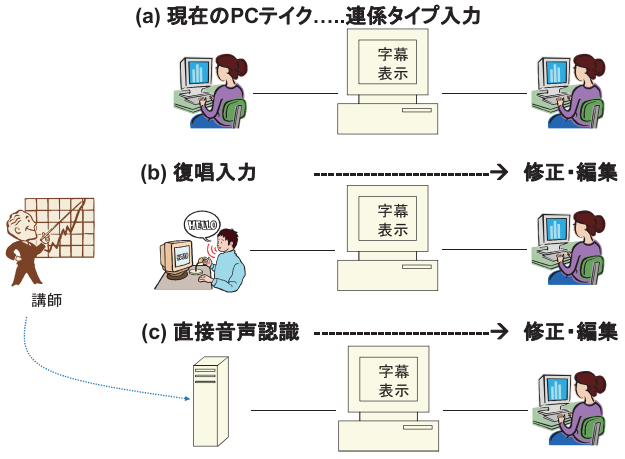

図７　リアルタイム字幕付与（ノートテイク）の方式

る。伊福部らは復唱入力方式を用いたシステム（図７(b)）を開発し，群馬大学などではこれを用いた実証実験を行っている（中野・牧原・金澤・中野・新井・黒木・井野・伊福部，2007）。訓練した復唱者によって90%程度の認識率を確保し，修正者も入れることによって，高い品質の字幕を実現している。ただしこの方式は，話し言葉を文法的かつ明瞭に発声し直す復唱者に大きな負担を課すことになり，２名交代での作業になる。その養成と確保は容易でない。

これに対して，著者らは講師の音声を直接認識する方式（図７(c)）を研究している。音声認識システムのモデルを講師に特化することによって認識精度を高めた上で，修正者による作業を経て字幕を作成する。事前に講演原稿がある場合には，これを元に単語辞書・言語モデルを構成することによって，90%程度の認識率を実現できる。毎年京都大学で開催している『聴覚障害者のための字幕付与技術』シンポジウムにおける著者の講演で，このシステムの実演を行っている。しかし原稿がない場合は，前述の通り，認識精度は60%～80%であるので，すべてを字幕として出すのは困難になる。京都大学工学部の講義で行った実験では，１名の修正者の作業で情報保障できたのは，講師の発話の30%～45%程度であった。ただし，それでも２名の手書きノートテイクに比べると２倍程度の情報量であった（勝丸・河原・秋田・森・山田，2009）。今後も，安定した音声認識精度を実現することと，どのように修正・提示を行うかについて，研究を進めていく必要がある。

6. おわりに

話し言葉の音声認識の最近の技術動向について解説を行った。議会の会議録作成に関しては実用的なレベルに到達したが，講演・講義に関してはまだまだ研究途上である。講演・講義は，話題や発話スタイルが多岐にわたるので，普遍的なモデルの構築が難しい。個々の講演・講義に効果的・効率的に音声認識システムを適応させる方法が鍵となっている。今後も，基礎研究ならびに実践的な応用の両面から進めていきたい。

謝辞

国会審議の音声認識システムならびに講演・講義の音声認識システムの研究開発に貢献頂きました秋田祐哉，三村正人両氏をはじめとする皆様に感謝します。

引用文献

秋田祐哉，三村正人，河原達也（2010）．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.

Akita, Y. and Kawahara, T.（2010）. Statistical transformation of language and pronunciation models for spontaneous speech recognition. IEEE Trans. Audio, Speech & Language Processing, Vol. 18, No. 6, pp. 1539-1549.

古井貞熙（2009）．人と対話するコンピュータを創っています　角川学芸出版.

Glass, J., Hazen, T. J., Cyphers, S., Malioutov, I., Huynh, D. and Barzilayet, R.（2007）. Recent progress in the MIT spoken lecture processing project. Proc. INTERSPEECH, pp. 2553-2556.

勝丸徳浩，河原達也，秋田祐哉，森信介，山田篤（2009）．講義音声認識に基づくノートテイクシステム　電子情報通信学会技術研究報告，SP2009-53, WIT2009-59.

河原達也（2006）．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.

河原達也（2012）．議会の会議録作成のための音声認識―衆議院のシステムの概要―　情報処理学会研究報告，SLP-93-5.

Kawahara, T.（2012）. Transcription system using automatic speech recognition for the Japanese Parliament（Diet）. Proc. AAAI/IAAI, pp. 2224-2228.

越仲孝文，江森正，大西祥史，北出祐，谷真宏，佐藤研治（2010）．法廷音声認識システムの開発―システム概要―　日本音響学会研究発表会講演論文集（春季），1-6-15.

前川喜久雄（2004）．『日本語話し言葉コーパス』の概観　国立国語研究所.

三村正人，秋田祐哉，河原達也（2011）．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.

中野聡子，牧原功，金澤貴之，中野泰志，新井哲也，黒木速人，井野秀一，伊福部達（2007）．音声認識技術を用いた聴覚障害者向け字幕提示システムの課題―話し言葉の性質が字幕の読みに与える影響―　電子情報通信学会論文誌，Vol. J90-D, No. 3, pp. 808-

814.
南條浩輝，秋田祐哉，河原達也（2005）．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　日本音響学会秋季研究発表会講演論文集，1-7-13.
Paul, M., Federico, M. and Stucker, S.（2010）. Overview of the IWSLT 2010 Evaluation Campaign. Proc. IWSLT, pp. 3-27.
Ramabhadran, B., Siohan, O. and Sethy, A.（2007）. The IBM 2007 Speech Transcription System for European Parliamentary Speeches. Proc. IEEE-ASRU.
Sakai, T. and Doshita, S.（1962）. The Phonetic Typewriter. Proc. IFIP Congress 62, pp. 445-450.
猿谷豊（2012）．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. 55, No. 6, pp. 392-399.
佐藤庄衛（2012）．音声認識を用いた生放送番組への字幕付与　メディア教育研究，Vol. 9, No. 1, S9-S18
Van Thong, J-M., Moreno, P. J., Logan, B., Fidler, B., Maffey, K., Moores, M.（2002）. SpeechBot: An experimental speech-based search engine for multimedia content on the web. IEEE Trans. Multimedia, Vol. 4, No. 1, pp. 88-96.
山本幹雄（2011）．広島大学における音声認識を活用した教育支援の取組　聴覚障害者のための字幕付与技術シンポジウム予稿集，pp. 9-15.
山崎恵喜（2006）．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.
吉川あゆみ，太田晴康，白澤麻弓（2001）．大学ノートテイク入門　人間社.

河原（かわはら）　達也（たつや）
2003年から京都大学学術情報メディアセンター／情報学研究科教授。音声言語処理，特に音声認識及び対話システムに関する研究に従事。主著に，「音声認識システム」「音声対話システム」（いずれもオーム社）。IEEE，情報処理学会，日本音響学会，電子情報通信学会，人工知能学会，言語処理学会各会員。

Recent Progress of Spontaneous Speech Recognition —Deployment in Parliament and Applications to Lectures—

Tatsuya Kawahara[1)]

In the past decade, there has been significant progress in the speech recognition technology. It has also been studied with regard to public speaking such as lectures and Parliamentary meetings, and several systems have been deployed in practical applications. This article describes the recent trend of the technology and applications including those to captioning and note-taking of lectures.

Keywords
Speech Recognition, Spontaneous Speech, Captioning, Note-taking, Meeting Record

[1)] Kyoto University