音声認識結果の有用性の自動判定に基づく講義のリアルタイム字幕付与システム

桑原 暢弘[1]　秋田 祐哉[1,2]　河原 達也[1,2]

概要：大学講義などの場面における情報保障として，音声認識を用いたリアルタイムの字幕やノートテイクが検討されている．しかし，認識誤りや話し言葉表現が含まれる音声認識結果を逐次的に編集して出力すると，提示に遅延が発生することは避けられない．そこで本研究では，効率的な編集・提示のために，字幕としての有用性の観点から音声認識結果を自動的に分類する手法，および自動分類に基づく字幕提示法を提案する．本研究では，構文的な正しさ・音声認識誤りの有無・話し言葉の冗長な表現の有無という点から有用性を定義する．これに基づき，ルールと機械学習を用いて，音声認識結果を「有効入力」・「無効入力」・「要チェック」の 3 種類に分類する．字幕の提示の際は，この自動分類結果をもとに，有効入力は速やかに提示し，要チェック箇所のみを人手でチェック・修正して表示する．本研究では，これらの手法からなるリアルタイム字幕付与システムを実際の講義において試行した．

Real-time Lecture Captioning System using Automatic Classification of Usability of ASR Result

Nobuhiro Kuwahara[1]　Yuya Akita[1,2]　Tatsuya Kawahara[1,2]

Abstract: As a support to hearing-impaired students attending classes, real-time captioning and note-taking using automatic speech recognition (ASR) have been investigated. However, even with ASR, editing by hand is needed to correct recognition errors and redundant spoken expressions in ASR results, and thus it often leads to delay in presenting captions. For efficient edit and quick presentation, we propose automatic classification of ASR results in terms of usability as caption, and then a presentation method based on the classification. In this study, we define the usability by syntactic correctness, errors and redundant spoken expressions in ASR results. Based on this definition, each unit of ASR results is classified into "valid," "invalid" or "to be checked," using hand-crafted rules and a machine learning framework. When presenting captions, valid input is presented promptly, and then checked ASR results are additionally provided after manual edit. We developed a real-time captioning system by combining the automatic classification method and the presentation method, and conducted a trial of this system in a university lecture.

1. はじめに

大学などの高等教育機関を対象とした調査によれば，聴覚に障がいを持つ学生は年々増えてきており，2012 年度にはおよそ 1,500 名と報告されている [1]．このような聴覚障がい学生を支援するため，講義などの場で音情報を視覚情報に変換して伝える情報保障の取り組みがなされている．大学の講義における情報保障としては，手書きのノートテイクやパソコンでタイプする PC テイクが一般的に行われているが，すべてを文字化することは困難で，ノートテイクの場合では発話全体の 2 割程度しか伝達できないのが現状である [2]．また，大学の講義は専門用語が多いことから，専門分野が同一のノートテイカーでないと聞き取り自体が難しく，ノートテイカーの養成や確保は容易ではない．

これに対して，音声認識をノートテイクに活用する試み [3][4] がなされており，我々も開発に取り組んでいる．

[1] 京都大学 情報学研究科
School of Informatics, Kyoto University
[2] 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

音声認識は人手と比べてはるかに高速であり，すべての発話を書き起こして出力できる．ただし誤認識や話し言葉表現の編集が必須であり，認識結果を逐次的に編集して出力することから提示に遅延が発生する．

そこで本研究では，音声認識結果を効率的に編集・提示するために，認識結果を有用性の観点から自動的に分類する手法を提案する．これは提示の可否や修正の要否で認識結果を分類するもので，そのまま提示できるものは速やかに出力し，修正の必要な認識結果のみに作業を限定することで遅延の削減を図る．

以降では，まず提案する自動分類手法について述べ，分類結果に基づく字幕提示法について検討する．そして，これらを用いて構成されるリアルタイム字幕付与システムについて，実際の講義における試行の結果とともに示す．

2. 音声認識結果における有用性

音声認識結果の有用性の指標として，認識結果やその中の単語に信頼度を付与することが一般的に行われている．たとえば，文仮説に出現する単語の事後確率をもとに信頼度を計算する手法 [5] などである．音声認識誤りを検出・訂正する研究 [6], [7] も行われており，近年では識別的なモデルを用いた手法 [8], [9], [10] も研究されている．

これらは「認識結果が正しいか否か」という観点から判定を行っているのに対して，本研究では「字幕としての有用性」の観点から分類を行う．すなわち，理解の妨げとならないような認識誤りは問題としないが，話し言葉の冗長な言い回しなどは正しく認識できたとしても棄却する．このような研究として，たとえば音声による質問応答システムにおける有効入力と無効入力の分類手法 [11] が提案されている．しかし本研究の場合は，単に有効（受理）・無効（棄却）というだけでなく，人手でチェックを行うかどうかも判断する必要がある．また，文献 [11] では質問応答システムに入力する音声認識結果を判定しているのに対して，本研究ではシステムの出力について判定している．このため，キーワードが正しいといった点だけでなく，出力全体が構文的に正しいかどうかも考慮する必要がある．

3. 音声認識結果の自動分類手法

本研究では，形態的・音響的な特徴を用いて，音声認識結果を「有効入力」・「無効入力」・「要チェック」のいずれかに分類する．自動分類する単位として，人間にとって直感的でわかりやすく処理しやすい文節を採用する．分類の手順としては，あらかじめ自動的に文節にまとめあげた音声認識結果に対して，まず構文的な情報をもとにルールによる分類を試みる．音声認識誤りに起因するものなど，ルールで決定できないものについてはさらに CRF（Conditional Random Fields）による分類を行い，最終的な結果を決定する．以降ではそれぞれのステップについて述べる．

3.1 文節へのまとめあげ

文節へのまとめあげは係り受け解析器で行うことができる．しかし，Cabocha[12] や JUMAN/KNP[13] 等の係り受け解析器は書き言葉を対象にしており，話し言葉である上に認識誤りが含まれる音声認識結果に対しては文節へのまとめあげ精度が低下する．

西光ら [14] は，機械学習の 1 つである SVM（Support Vector Machines）を用いて，『日本語話し言葉コーパス』（CSJ）から学習したモデルにより文節へのまとめあげを行っており，音声認識誤りのある文に対しても比較的頑健に機能すると報告している．本研究ではこの手法を用いる．

3.2 分類の定義

本研究では，「有効入力」「無効入力」「要チェック」の 3 種類への分類を行う．これらの分類にあたって考慮すべき要因として次の 3 点がある．

(a) 構文的な正しさ：文節へのまとめあげ結果が，文法的に適切であるか．

(b) 内容語の認識の正しさ：音声認識結果の内容語（名詞・動詞・形容詞・副詞・複合名詞/動詞）がすべて正しく認識されているか．

(c) 表現の冗長さ：言い直し・言い淀み・話者特有の口癖・呼応の副詞のように，字幕として冗長・不要な音声認識結果でないか．

内容語に認識誤りがなく，文節へのまとめあげが適切であり，かつ冗長語が含まれないものを「有効入力」と定義する．文節へのまとめあげが適切でない，または認識誤りが含まれるものは「要チェック」，冗長語が含まれるものは「無効入力」と定義する．これら 3 種類の分類は，次の通り (a)・(b)・(c) の要因の組み合わせで表すことができる．

有効入力　字幕に提示すべき：(a) かつ (b) かつ (c)

無効入力　字幕に提示すべきでない：($\bar{c}$)

要チェック　確認・修正の上，字幕に提示すべき：上記以外

3.3 ルールと CRF による分類

(a) に関しては，文法的なルールを定めることができる．一方，(b)・(c) については網羅的に記述することは困難であるため，CRF で判定する．本研究では，まずルールによる判定を行い，次いで CRF による判定を行う．

ルールによる判定では，音声認識結果の各単語の品詞情報を用いて，正規表現 (1) に従うかどうかで分類を行う．

$$\text{付属語}^{*}\ \text{接頭辞}^{*}\ \text{自立語}^{+}\ \text{接尾辞}^{*}\ \text{付属語}^{*} \qquad (1)$$

‘$*$’ は 0 回以上の繰り返し，‘$+$’ は 1 回以上の繰り返しを表す．正規表現 (1) に従わない文は文節として正しくない構造であるから，要チェックに分類する．ただし付属語のみの文は字幕としては不要と考えられるため，無効入力と分類する．

表 1　自動分類の評価セット

講演	単語正解精度（%）	総文節数	ラベルの割合（%）		
			有効入力	無効入力	要チェック
話者 A	76.5	1,620	64.0	7.7	28.3
話者 B	77.9	1,537	70.3	5.3	24.3
話者 C	81.6	1,508	73.9	6.0	20.1

表 2　素性の組み合わせと分類精度

番号	素性						分類精度（%）
	文節	読み	品詞情報	信頼度スコア	内容語の数	ポーズの有無	
(1)	○	×	×	×	×	×	55.6
(2)	○	○	×	×	×	×	55.8
(3)	○	○	○	×	×	×	75.3
(4)	○	○	○	○	×	×	79.7
(5)	○	○	○	○	○	×	79.5
(6)	○	○	○	○	○	○	79.6

正規表現 (1) に従う文はさらに CRF により分類する．ここでは，3 種類のラベルを付与する系列ラベリング問題として分類を考え，CRF でモデル化する．CRF で用いる素性は，文節の表層表現と読み，音声認識の信頼度スコア，ポーズの有無，品詞情報，文節に含まれる内容語の数である．このうち音声認識の信頼度スコアには，まとめあげた文節に含まれる内容語の信頼度スコアの平均値を用いる．CRF の実装としては CRF++[*1]を利用する．

3.4　評価実験

本研究では，京都大学 iPS 細胞研究所の公開シンポジウムにおける 3 件の講演を対象として有用性判定の評価を行った．評価で用いた講演のデータを**表 1** に示す．各講演の音声認識結果を文節へまとめあげ，3.2 節の定義に従って人手で「有効入力」「無効入力」「要チェック」の 3 種類にラベル付けしたものを評価に利用している．

まず，各素性の有効性を明らかにするため，素性の組み合わせを変えて評価を行った．実験結果を**表 2** に示す．ここでは講演単位でオープンにした 3 分割の交差検定を行っている．評価指標としては，正解した文節数を評価文節数で割った分類精度を用いる．表 2 より，(4) の文節，読み，品詞情報，信頼度スコアを素性として用いた場合に最も高い分類精度 79.7%が得られた．

次に，各ラベルごとの結果について考察する．**表 3** に，素性として (4) の文節，読み，品詞情報，信頼度スコアを用いた場合の分類結果を示す．有効入力が要チェックに，無効入力が有効入力に誤分類されている割合が多いが，これらは実際には大きな問題とはならない．一方，要チェックと有効入力の混同が多いが，要チェックを有効入力と誤分類したものは，修正・確認をすべき認識結果をそのまま提示することになり，改善の必要がある．これらの分類においては (*b*) の判定が重要であるから，(*b*) に特化した識別器を大規模に学習することで分類精度が改善できると考えられる．無効入力については他よりも再現率が低いが，無効入力の事例が少ないため，事例を増やせば分類精度が向上する可能性がある．

[*1] http://code.google.com/p/crfpp/

表 3　分類結果（素性：文節・読み・品詞情報・信頼度スコア）

判定 / 正解	有効入力	無効入力	要チェック	合計	割合（%）	再現率（%）
有効入力	2,892	14	325	3,231	69.3	89.5
無効入力	114	141	43	298	6.4	47.3
要チェック	443	10	683	1,136	24.4	60.1
合計	3,449	165	1,051	4,665		

表 4　話者ごとの分類精度

話者	単語正解精度（%）	分類精度（%）	
		話者オープン	話者依存
話者 A	76.5	77.8	77.7
話者 B	77.9	77.8	78.8
話者 C	81.6	83.6	82.8
平均	78.7	79.7	79.8

これまでに述べた評価では，話者に関してオープンなテストとなっているが，逆に話者に依存したモデルとすることも考えられる．そこで，話者ごとに 10 分割の交差検定による評価を行った．この場合，素性として文節，読み，品詞情報，内容語の数，ポーズの有無，信頼度スコアをすべて用いた場合に高い分類精度が得られた．それぞれの話者における単語正解精度と分類精度を**表 4** に示す．表 4 より，単語正解精度と分類精度とに相関がみられる．3 者の分類精度の平均は 79.8%であり，話者オープンの分類精度 79.7%とほとんど変わらない．これによって，話者オープンのモデルでも十分であることが示された．

4.　自動分類に基づく字幕提示法

音声認識を用いたリアルタイム字幕（たとえば [3]）では，入力を逐次的に編集して出力する．これに対して，本研究で提案する自動分類を用いると，有効入力について即時に出力が可能である．すなわち，編集を待って順次出力するのではなく，有効入力は速やかに出力し，編集した部分はあとから反映させることができる．本節では，有用性判定に基づく提示手法と従来の提示手法について比較し，聴覚障がい者の立場からの評価について述べる．

4.1　提示方法

本研究では 3 種類の提示方法を比較する．字幕の文字表示スタイルはすべての手法に共通とし，黒い背景に原則として白い文字を用いて，フォントは MS ゴシック・24 ポイントと設定した．

表 5　字幕提示システムの評価に用いる講演

字幕提示法	講演データ	単語正解精度 (%)
手法 1	話者 A	77.9
手法 2	話者 B	76.5
手法 3	話者 C	81.6

手法 1 は音声認識結果を確認・修正した後の字幕のみを表示する手法で，従来から用いられているものである．リアルタイム性は失われるが，正確な情報のみを提示できる．

手法 2 は音声認識結果をまず灰色で表示し，確認・修正した字幕を白色で上書きして字幕を表示する手法である．本手法は，リアルタイム性を重視した字幕と，正確性を重視した字幕の両方を提示できる．しかし 2 種類の字幕を見るため，身体的な負担が増すおそれがある．

手法 3 は本研究で提案する手法で，自動分類の結果を反映させた音声認識字幕をまず表示し，要チェック箇所を修正した後に上書きして字幕を提示する．具体的には，有効入力は字幕をそのまま提示し，無効入力は提示せず，要チェックは文字数だけ '-' を表示する．そして，修正結果は赤文字で上書きする．本手法は，正しい認識結果を即座に提示し，音声認識誤りを含むものは提示せず，人手で修正してから提示する．リアルタイム性と正確性のバランスを考慮した手法である．

4.2　評価実験

提示法の評価は実際の講義で行うことが望ましいが，聴覚障がいの被験者の確保，講義内容の統一などの点から，実際の講義での評価は困難である．そこで本研究では，講演の動画を用いた，シミュレーションによる実験を実施した．本実験では音声認識文と修正・確認後の文を事前に作成しておき，講演の動画と同期させて提示を行う．音声認識結果を提示するタイミングは，音声認識器の出力時刻と同一になるように設定している．また，音声認識結果を確認する時間を 1 文節あたり 500ms とし，修正する時間を一文字あたり 500ms と設定した．これは修正・確認作業を一人で行うことを想定し，テイカーに要求されるタイピング速度と日本人の平均的な読む速度に合わせたものである．**表 5** に実験で使用するデータを示す．各データは 10 分程度とし，被験者ごとに提示法の順番を変えて実験を行った．被験者は現在高等教育機関に通っている，あるいは最近卒業した 18 歳〜26 歳の男性 3 名，女性 3 名（計 6 名）である．いずれもふだんからノートテイク・PC テイクを受けている，または受けた経験がある人である．被験者の前に字幕提示用としてノートパソコンを置き，4 メートル先のスクリーンに動画を映して評価実験を行った．

実際の講義に近づけるためには，被験者が内容理解に努めるような動機付けが必要である．そこで被験者に，動画内容を理解したかどうかの確認テスト（各方式 5 問ずつ）を行った．たとえば，「パーキンソン病は，どんな症状が出るのか理解できましたか?」のような質問に対して，Yes/No で回答してもらう．各方式ごとの 6 名の正解率の平均は，手法 1 が 93.3%，手法 2 が 76.7%，手法 3 が 80.0%であった．手法 1 は常に正しい情報だけ表示され，かつ被験者にとって慣れている方式であるため，最も高い正解率が得られたと考えられる．

表 6　順位付け結果

	手法 1	手法 2	手法 3
被験者 A	3	2	1
被験者 B	1	2	3
被験者 C	3	2	1
被験者 D	1	3	2
被験者 E	2	3	1
被験者 F	1	2	3
平均	1.83	2.33	1.83

被験者に，3 つの字幕提示システムのうち実際に使うとすればどのシステムがよいか順位付けしてもらった結果を**表 6** に示す．被験者の支持は手法 1 と手法 3 に分かれる結果となった．このうち手法 3（提案手法）については，「ところどころ抜け落ちていても全体的な意味は理解できる」「リアルタイムで字幕を提示してくれるので助かる」とのコメントが実験後のヒアリングで得られた．一方，手法 1 に関しては，被験者 B はふだんから PC テイクを受けているため，バイアスがかかったと考えられる．また被験者 F は聴覚フィードバックが全く得られないため，手法 3 が最も使いづらいシステムになったのではないかと考えられる．これらの結果から，本研究で提案する手法 3 は一定の支持が得られたといえる．

5.　リアルタイム字幕付与システム

5.1　システムの構成

これまでに述べた自動分類手法と提示手法を用いて，リアルタイムの字幕付与システムを構築した．このシステムの構成を**図 1** に示す．まず，ワイヤレスマイクを通じて講師の音声を入力し，リアルタイムで音声認識を行う．次に，音声認識で得られた形態素列から「あのー」「えー」のようなフィラーを削除し，形態素列を自動的に文節へまとめあげる．そして，3.3 節で述べた自動分類器により，文節を「有効入力」「無効入力」「要チェック」の 3 種類に分類する．有効入力は修正せずにそのまま出力し，無効入力は出力せず棄却する．要チェック入力は修正者による確認・編集を行うが，これにはパソコン要約筆記に一般的に用いられる IPtalk[*2]を用いる．字幕の提示装置については，講師や黒板と字幕との間の視線移動を削減するため，透過型の表示装置であるプロンプターを机上に設置して利用する．

[*2] http://www.geocities.jp/shigeaki_kurita/

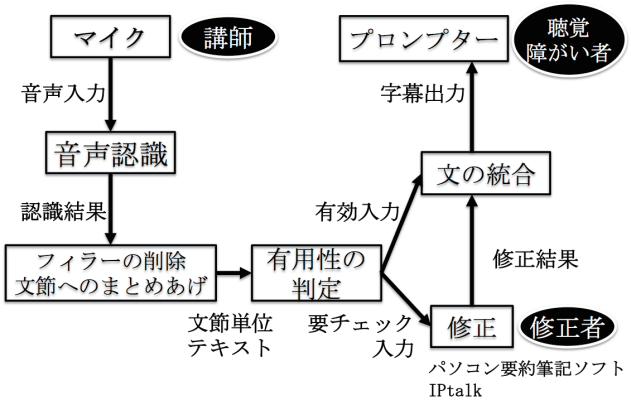

図 1　リアルタイム字幕付与システムの構成

5.2　大学講義における試行

本研究のリアルタイム字幕付与システムを，京都大学情報学研究科における実際の講義で試行した．本講義を履修している聴覚障がい学生（1 名）に対して，比較のため，2 回の講義でそれぞれ異なる字幕付与システムにより情報保障を行った．1 回目では従来法，すなわち 4.1 節で述べた手法 1 を実施し，音声認識結果のすべてに対して人手によるチェック・修正を行って字幕として提示した．2 回目では提案システム（手法 3）で字幕を作成して提示した．

音声認識については，デコーダとして Julius 4.2.3 を利用し，リアルタイムで認識できるようにデコーディングのパラメータをあらかじめ調整した．音響モデルには CSJ の学会講演モデル [15] を用い，声道長正規化（VTLN）に加えて，当該講師の過去の音声を用いた MLLR 話者適応を行っている．言語モデルについても，CSJ の講演書き起こしに当該講師の過去の講義スライドテキストや書き起こしを混合して，話題に適応したモデルを構築した．この音声認識器は 1 回目・2 回目に共通して用いられている．

講義ののち，京都大学の障害学生支援ルームを通じて当該学生にヒアリングを実施した．まず字幕のリアルタイム性について，従来システム（1 回目）より提案システム（2 回目）のほうが速く感じたとの回答が得られ，提案システムで期待したリアルタイム性の向上が確かめられた．ただし，いずれのシステムにおいても字幕の精度が低く，被験者にとって大きな負担となった．音声認識の精度（文字正解精度）は，1 回目が 58.1%，2 回目が 61.8%で，いずれも十分な精度ではなかった．これにより自動分類の精度が低下し，また作業者による修正量が増大したことが字幕の精度低下の要因である．なお，システムの試行に先だって行ったリハーサル（文字正解精度 69.8%）ではスムーズに修正できたとのコメントが修正者から得られており，本システムの運用には 70%程度の文字正解精度が必要であることが示唆された．このほか，プロンプターに関して，字幕を見ながら講師の顔が見えるので状況が判断しやすいとの回答が得られ，本研究で想定した効果が得られたといえる．

6.　おわりに

本稿では，字幕としての有用性の観点から音声認識結果を自動で分類する手法と，この分類に基づいた提示手法からなる，講義のためのリアルタイム字幕付与システムを提案した．この字幕提示手法について，聴覚障がい者による評価を実施したところ一定の支持が得られた．さらに，システムを実際の大学講義で試行し，字幕のリアルタイム性が向上したことを確認した．

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．字幕付与システムの試行にご協力いただきました，京都大学情報学研究科准教授 山肩洋子先生に感謝いたします．

参考文献

[1] 日本学生支援機構: “大学、短期大学及び高等専門学校における障害のある学生の修学支援に関する実態調査”，2013.
[2] 斎藤佐和，白澤真弓，徳田克己： “聴覚障害学生サポートガイドブック”，日本医療企画，2002.
[3] T.Kawahara, et al.: “Classroom Note-taking System for Hearing Impaired Students using Automatic Speech Recognition Adapted to Lectures”, In *Proc. Interspeech*, pp.626–629, 2010.
[4] P.Cerva, et al.: “Real-time Lecture Transcription using ASR for Czech Hearing Impaired or Deaf Students”, In *Proc. Interspeech*, 2012.
[5] F.Wessel, et al.: “Confidence Measures for Large Vocabulary Continuous Speech Recognition”, *IEEE Trans. Speech & Audio Process.*, Vol.9, No.3, pp.288–298, 2001.
[6] Z.Zhou, H.Meng and W.K.Lo: “A Multi-pass Error Detection and Correction Framework for Mandarin LVCSR”, In *Proc. Interspeech*, pp.1646–1649, 2006.
[7] A.Allauzen: “Error Detection in Confusion Network”, In *Proc. Interspeech*, pp.1749–1752, 2007.
[8] 中谷良平，滝口哲也，有木康雄: “CRF と Confusion Network を用いた音声認識誤り訂正”，日本音響学会春季研究発表会講演論文集，2-P-59(a)，2011.
[9] Z.Zhou, et al.: “A Comparative Study of Discriminative Methods for Reranking LVCSR N-best Hypotheses in Domain Adaptation and Generalization”, In *Proc. ICASSP*, Vol.1, pp.141–144, 2006.
[10] G.Kurata, N.Itoh and M.Nishimura: “Traning of Error-corrective Model for ASR without Using Audio Data”, In *Proc. ICASSP*, pp.5576–5579, 2011.
[11] H.Majima, et al.: “Spoken Inquiry Discrimination using Bag-of-words for Speech-oriented Guidance System”, In *Proc. Interspeech*, 2012.
[12] T.Kudo and Y.Matsumoto: “Japanese Dependency Analysis using Cascaded Chunking”, In *Proc. CoNLL*, pp.63–69, 2002.
[13] 笹野遼平, 河原大輔, 黒橋禎夫, 奥村学: “構文・述語項構造解析システム KNP の解析の流れと特徴”, 言語処理学会第 19 回年次大会発表論文集, pp.110–113, 2013.
[14] 西光雅弘, 秋田祐哉, 高梨克也, 尾嶋憲治, 河原達也: “局所的な係り受けの情報を用いた話し言葉の節・文境界の推定”，情報処理学会論文誌，Vol.50，No.2，pp.544–552，2009.
[15] 三村正人, 河原達也: “大学講義の音声認識のための音響・言語モデル適応に関する検討”, 日本音響学会秋季研究発表会講演論文集, 3-P-6, 2011.