Vol.2016-SLP-112 No.12
2016/7/30

講演のリアルタイム字幕付与のための音声認識結果の簡約

大田 健翔[1]　秋田 祐哉[2]　河原 達也[1]

概要：本研究では，聴覚障がい者への情報保障のために，講演に対する音声認識を用いたリアルタイムの字幕付与を扱う．話し言葉を音声認識で書き起こす際には，冗長な語句も認識結果として出力されるため文字数が増えて読みにくくなる．そこで本研究では，文意を保存しつつ冗長な語句を削減する簡約処理を検討する．具体的には，講演内容を理解するにあたって必要な単語（内容語）とそうでない単語（付属語）に分類し，原則として後者を削除し前者のみを残して字幕として提示する．この原則にあてはまらないものがあるので，内容語で削除するものをアノテーション頻度の比率に基づいて決定し，付属語で復元するものをアノテーション頻度の比率，N-gram による言語尤度比較，機械学習を用いる方法で決定する．講演音声の書き起こしに対して簡約処理を行った結果，正解率 78%・圧縮率 64%で文を圧縮することができた．

Condensation of Speech Recognition Results for Real-Time Lecture Captioning

Kensho OTA[1]　Yuya AKITA[2]　Tatsuya Kawahara[1]

Abstract: We have been investigating a real-time captioning framework using automatic speech recognition (ASR) technology for hearing-impaired audience. Since an ASR system transcribes all of speech input, including redundant spoken expressions, resulting captions are very long and thus hard to read and understand. To solve this problem, we propose a "condensation" method, which reduces unnecessary expressions in ASR results as much as possible while retaining key meaning of the utterances. Specifically, each word in ASR results is classified into a content word or a dependent word. Basically, the latter is deleted, while the former is retained for captions. However, there are exceptions in this principle, thus we further introduce refinement process. Redundant content words to be deleted are determined using occurrence counts in annotated training data. On the other hand, for recovery of dependent words, we investigate three methods: occurrence counts in annotated training data, linguistic likelihood measure calculated by an N-gram language model, and a machine learning framework. In an experiment over real lecture transcriptions, word-based compression rate of 64% and accuracy of 78% was obtained.

1. はじめに

近年，障がい者の社会参画機会を均等化する機運が高まっており，2016 年度には障害者差別解消法 [*1] が施行された．この法律に則り，障がいによる社会的障壁除去のための「必要かつ合理的な配慮」が行政機関等および事業者の努力義務（国及び地方公共団体は義務）となっている．このうち，聴覚障がいへの配慮としては，音に代わる手段で情報保障を提供する必要がある．本研究ではこのような背景を受けて，講演の字幕提供により聴覚障がい者への情報保障を支援することを考える．

現在，講演や講義などでは文字通訳が一般的に行われている．講師の発話を文字にする方法としては，従来より，作業者が手書きで要約筆記を行ったり，複数人が連携して PC でひとつなぎの字幕を作る PC テイクが行われている [1][2]．しかし，要約筆記では書き取る速度よりも講師の話す速度の方が速いために，講演内容を網羅することは難

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

[*1] http://www8.cao.go.jp/shougai/suishin/sabekai.html

しい．PC テイクでは発言内容を文字にタイプするが，講演を安定して書き起こすために必要な人を確保することが困難である．

そこで本研究では，講師の発話内容を音声認識で自動的に文字に書き起こし，視覚情報として聴覚障がい者へ字幕を提供する方式を考える．これまで音声認識を用いた字幕付与手法およびシステムはいくつか提案されている [3][4]．音声認識を字幕生成に用いるメリットは，内容を網羅した字幕作成を少人数で行える点である．また，音声認識器のモデルを講演内容に適応すれば，講演で扱う専門用語の認識も可能となる．

ただし，話し言葉を音声認識で書き起こす際には，冗長な語句まで認識結果として出力されるため文字数が増え読みにくくなる．現行の字幕付与システムによる運用では作業者は誤認識の修正が作業の限度であり，冗長性の削減をする余裕まではない．そこで本研究では，文意を保存したまま冗長な語句をできるだけ自動的に削減する簡約処理を行うことで読みやすさの改善を図る．具体的には，講演内容を理解するにあたって必要な語（内容語）とそうでない語（付属語）に分類し，後者を削除し前者のみを残して字幕として提示することを基本とする．ただし，削除すべき内容語，削除すべきでない付属語という例外もあるため，これらの自動検出を行う．この処理によって，内容語と必要最低限の付属語で字幕が構成され，視聴者にとって読みやすいものとなることが期待される．

2. 話し言葉の冗長表現の削減

音声認識は，原則としてすべての発話内容を書き起こす．そのため，講演のような自然発話の場合，フィラーなどの非流暢表現や冗長な文末表現なども書き起こされてしまう．これらをすべてそのまま表示すると，かえって読みにくさにつながる．この例を図 1（A）に示す．

本研究では冗長な表現の削減，すなわち，発話の理解の上で必要性が小さい部分をできるだけ削除することを目指す．これにより，作業者が修正すべき音声認識誤りの箇所も減らすことが期待できる．以下の節では，話し言葉の書き起こしに対してどのような目的でどの冗長表現を削減するかという観点から既存手法を分類して述べる．

2.1 話し言葉の整形

整形は，話し言葉を書き言葉に近い形で人が読めるように，話し言葉特有の表現を処理するものである．対象として，講義録への適用 [5], [6] や議事録作成 [7] の例が報告されている．図 1（A）の書き起こしを整形処理すると図 1（B）のように文章が整えられる．整形は表層的な表現の操作にとどまり，文意の理解に重要ではない部分も残ることになる．この点で，図 1（C）に示す要約（後述）とは異なる．例えば「個人的に」といった表現は整形においては残るが，要約では削除される．

（A）書き起こし

えーっと 私 は あのー 個人 的 に あの 海外 に 興味 元々 興味 が ありまして えーと 社会人 に なって 自分 で 旅費 を 払える ように なって からは そう ですね 年 に 多くて 三回 えー 少ない 時 でも 一回 は あのー 海外 旅行 を して います

（B）整形した文

私 は 個人 的 に 海外 に 元々 興味 が ありまして 社会人 に なって 自分 で 旅費 を 払える ように なって からは 年 に 多くて 三回 少ない 時 でも 一回 は 海外 旅行 を して います

（C）要約した文

私 は 海外 に 興味 が あり 社会人 に なって 旅費 を 払える ように なって からは 年 に 三回 少ない 時 でも 一回 は 海外 旅行 を します

（D）簡約した文

海外 に 元々 興味
社会人 に なる
自分 で 旅費 払える
年 多く て 三回 少ない 時 一回 旅行

図 1 講演の書き起こし・整形・要約・簡約の例

2.2 話し言葉の要約

要約とは単一ないし複数の文書から重要な内容を抽出，もしくは冗長な部分を削除して新たな文章を作ることを指す．図 1（A）の書き起こしを要約処理すると図 1（C）のように文章がまとめられる．

話し言葉の要約においては，重要文抽出よりも，文圧縮の手法が望ましい．その理由は，話し言葉には文という明確な概念がないため抽出する単位を決定することが難しいことや，文抽出では文脈の整合性を意識することから文章の大域的な構造を見る必要があり，リアルタイム性の観点から見て逐次処理には向かないことが挙げられる．文圧縮の手法としては，構文解析木を操作することによって冗長な単語を削除する手法 [8] や，構文情報に基づく素性から単語を削除するかどうかを決定する分類問題 [9] としたり，膨大な原文と圧縮文からなる要約ペアを元に LSTM でモデル化する手法 [10] がある．しかし，これらの技術をそのまま音声認識結果に適用することは難しいと考えられる．その理由は，自然発話や認識誤りにより構文解析の精度が低下することである．また，要約ペアを大量に作ることが困難である．

講演・講義音声認識結果を対象としたリアルタイム圧縮型の要約手法として，Ohno ら [11] は話し言葉を漸近的に係り受け解析しながら要約単位を決定し，係る単語を修飾語とみなして，削除するという手法を提案している．しかし，係り受けが有用なのは，認識誤りが非常に少ない場合

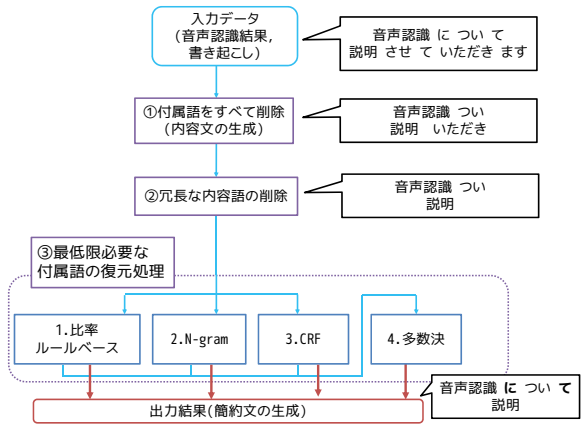

図 2 簡約処理のフロー

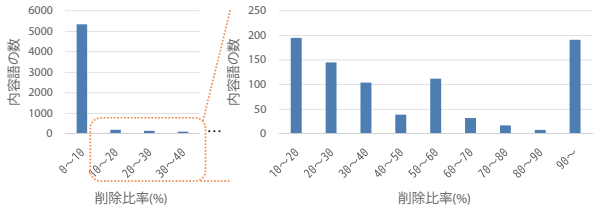

図 3 削除された内容語の頻度比率におけるヒストグラム（右のグラフは左のグラフの 10%以降を拡大したもの）

であり，この研究でも実際の音声認識結果では要約を行っていない．一方 Furui ら [12] は認識結果の単語に重要度を付与して，音声認識の信頼度を考慮しながら言語スコアの高いパスを動的計画法により探索し，最適なものを要約結果とする手法を提案している．これにより，誤認識を抑えた要約が可能となる．Ohno らと Furui らの研究はいずれも文法的正しさを重視したものとなっており，冗長性が依然として残っている．反面，内容語の一部も削除される．この発展として，N-gram 確率や音声認識の信頼度・N-best 仮説に基づく最適化問題として文圧縮を解く枠組みが提案されているが [13]，必要な計算量が大きい．

本研究ではこれらの課題を解決するために，講演の内容として必要な語のみを字幕として出力する方法を検討する．すなわち，必要な内容語を残して，助詞などの付属語に関して意味が通じる範囲で全て削除することとする．処理後の文章は非文になることもあるが，非文はリアルタイムの字幕においては許容されると考えられる．このように，「必要な情報は網羅しながらも，提示する文字数を最小限まで圧縮する処理」のことを本研究では**簡約**と呼ぶ．図 1（A）の書き起こしを簡約処理すると図 1（D）のように文字数が大幅に少なくなる．

3. 提案手法 —話し言葉の簡約—

3.1 簡約処理のフロー

提案手法では，原則として内容語（名詞，動詞，形容詞，副詞，接頭辞，接尾辞，記号）を字幕として表示する．ただし，冗長と判断される内容語は削除し，意味が通じない場合は例外として付属語（内容語以外）を字幕に含める．

90%以上削除された
内容語とその比率 (%)

やっと: 100.0
存じ: 100.0
おく: 100.0
つき: 100.0
みる: 100.0

50%～70%で削除された
内容語とその比率 (%)

思う: 50.0
いこう: 52.94
申し上げ: 53.33
もう 67.44
思い: 69.79

図 4 アノテーションで削除された内容語の例

高頻度で復元される
内容語とその比率 (%)

まで: 75.76
られる: 76.62
や: 76.67
ば: 77.70
だけ: 78.79

低頻度で復元される
付属語とその比率 (%)

ます: 3.01
そして: 3.48
まし: 4.16
です: 4.89
で: 5.75

図 5 アノテーションで復元された付属語の例

例えば「性能 比較 明らか (で) あり ませ ん」では，付属語「で」を削除すると意味が通じにくいため復元する．

簡約処理した文（簡約文）の生成手順は以下のとおりである．フローを図 2 に示す．

- 手順 1: 音声認識結果や書き起こしの内容語以外を削除した文章（内容文）を作成する．
- 手順 2: 内容文からさらに冗長と判断される内容語を削除する．
- 手順 3: 上記で生成された文に対して内容理解に必要な付属語を復元する．

削除すべき冗長な内容語，および復元すべき付属語は以下の手順でアノテーションして学習する．

(1) 整形済みの講演書き起こしと，書き起こしの内容語以外を削除した文章（内容文）を用意する．
(2) アノテータが内容文を読み，不要な内容語，意味の通らないと判断した箇所に対して元の書き起こしを参照しながら付属語を補うアノテーションを行う．
(3) 内容文とアノテーションした文でアライメントを取り，削除すべき内容語，および復元すべき付属語を得る．

アノテーションの結果，内容語のうち削除された頻度の比率に関するヒストグラムを図 3 に示す．内容語の大半は削除されないので 0%に集中している．一方，90%以上で削除される内容語がかなりの割合であることがわかる．90%以上削除された内容語と 50%～70%で削除された内容語の例を図 4 に示す．高頻度に削除された内容語では，「思い （ます）」などの文末表現や「もう」などの話し癖の表現が多い．高頻度のものは機械的に削除し，低頻度のものは作業者が修正した方がよいと考えられる．

また，高頻度および低頻度に復元された付属語の例を図 5 に示す．高頻度に復元された付属語では，「や (and)」や「だけ (only)」など，副詞の意味合いをもつ助詞・助動詞が多い．低頻度に復元されたものは接続詞や文末表現が多い．

以上の手順で学習した単語を例外処理として登録し，不要な内容語の削除，および必要な付属語の復元を自動で行う．

3.2 冗長な内容語の削除

内容語の削除の可否は，アノテーションされた単語の出現比率を用いる．この比率のしきい値を θ_c とすると内容語 C の削除の可否は次のように決定される．

$$\begin{cases} \text{if } \dfrac{\text{内容語 } C \text{ が削除とアノテーションされた数}}{\text{ある内容語 } C \text{ の出現頻度}} \geq \theta_c \\ \quad \text{then 削除する} \\ \text{otherwise 削除しない} \end{cases}$$

図 3 で示した通り，ほぼ確実に削除してよい冗長な表現のみを削除する．その閾値を評価実験により決定する．

3.3 必要な付属語の復元

付属語の復元処理については下記の 4 種類のモデルを検討する．

- 手法 1: 頻度比率ルールベース
- 手法 2: N-gram による言語尤度比較
- 手法 3: 条件付き確率場 (CRF) による判定
- 手法 4: 上記 3 つの判定による多数決

手法 1 の頻度比率ルールベースは，内容語削除と同様に，アノテーションの比率によって復元するかどうかを決めるものである．比率のしきい値を θ_f とすると，ある付属語 F の復元処理は次式で決定される．

$$\begin{cases} \text{if } \dfrac{\text{付属語 F が復元とアノテーションされた数}}{\text{ある付属語 F の出現頻度}} \geq \theta_f \\ \quad \text{then 復元する} \\ \text{otherwise 復元しない} \end{cases}$$

手法 2 の N-gram による言語尤度比較は，内容語間に存在する付属語に対して，復元した場合としない場合とで単語 N-gram モデルにより言語尤度を求めて比較する方法である．例えば「C_1, F, C_2」のように内容語 C_1，C_2 の間に付属語 F があった場合，次式で復元の可否を定める．

$$\begin{cases} \text{if } \ p(F|C_1) * p(C_2|C_1, F) > p(C_2|C_1) \ \text{ then 復元する} \\ \text{otherwise 復元しない} \end{cases}$$

なお，付属語が文の終端の場合は無声区間を示す特殊記号<sil>を末尾に置き，言語尤度を計算する．内容語間に付属語が 2 つ以上続く場合は，単語列のべき集合を考え，すべての要素に対して言語尤度を計算し，最も言語尤度の高い単語列を選択する．本研究では「日本語話し言葉コーパス」（CSJ）の学会講演・模擬講演から単語 trigram モデルを構築して使用する．

手法 3 の CRF による判定では，アノテーションデータから，復元される付属語，復元されない付属語の 2 値分類器を CRF の枠組みで学習し，復元するかどうかを判定する．用いる入力素性は，注目している付属語，前の単語，後ろの単語，それぞれの単語の品詞である．前後の単語の範囲は 1 つから 3 つまで変化させて性能を評価する．ただし前後の単語には削除された内容語と復元されなかった付属語は除く．CRF のツールには CRF++[*2] を用いる．

手法 4 の多数決では，手法 1〜3 でそれぞれ独立に判定された結果をもとに，多数決で復元するかどうかを決定する．

表 1 正解と予測結果の対応関係

予測結果＼正解	正例	負例
正例	a	b
負例	c	d

4. 評価実験

内容語の削除と，付属語の復元に関して各手法の予測性能を評価した．評価には正例の予測精度だけでなく，負例の予測精度も調査する．また正例と負例の正解に着目した正解率も調査する．例えば，内容語の削除であれば，削除すべき内容語に対して正しく削除するよう予測できているかだけでなく，削除すべきでない内容語に対して削除を行わないという予測ができているかも調べる．

評価尺度の説明のために，正解と分類器の予測結果の対応関係を表 1 のように定義する．これにより正解率を次のように定める．

$$\text{正解率} = \frac{a+d}{a+b+c+d}$$

講演の書き起こしデータとして，CSJ の学会・模擬講演のコアから 50 件を選び，アノテーションを行った．1 つの講演に対して 1 名のアノテーターが作業を行った．評価の際は，10 分割の交差確認法を行い，上記評価尺度の平均値を算出した．

4.1 内容語の削除性能

まず，内容語の削除に関する正解率を図 6 に示す．横軸はしきい値 θ_c（3.2 項参照）を表す．θ_c を大きくするに従って，アノテーションで削除頻度の高い内容語のみが削除されるようになる．

注目する θ_c は，正解率が最大となる 40%，正例の適合率および負例の再現率が最大となる 70%，図 3 で高頻度と低頻度との境界となる 90%である．それぞれの θ_c で多数決の際にどのように影響するかは後述する．

4.2 付属語の復元性能

次に付属語の復元手法の性能に関して評価する．

まず，図 7 に頻度比率ルールベースにおける復元の正解率を示す．横軸は付属語の復元比率のしきい値 θ_f である．しきい値を大きくすると，アノテーションで復元頻度の高い付属語のみが復元されるようになる．およそ 50%付近で正解率が急激に変化することから，ここをしきい値として

[*2] https://taku910.github.io/crfpp/

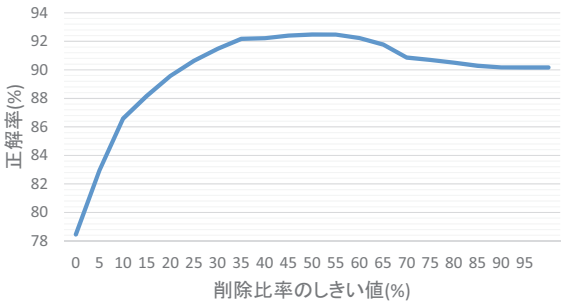

図 6 内容語削除の正解率

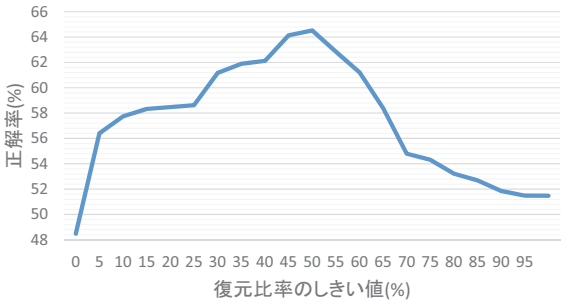

図 7 復元手法 1: 頻度比率ルールベースによる正解率

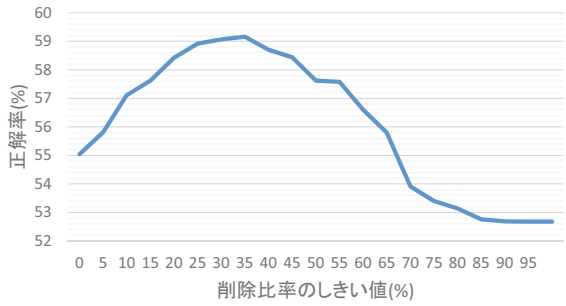

図 8 復元手法 2: N-gram による正解率

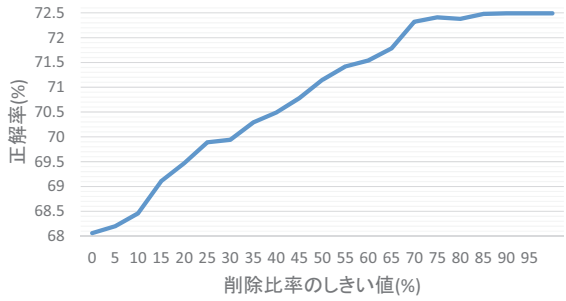

図 9 復元手法 3: 前後 2 単語を文脈とした CRF による正解率

定める．このとき正解率は最大の 64%となった．

次に，N-gram による復元の正解率を図 8 に示す．内容語の削除の影響を調べるために，このしきい値 θ_c を変化させて性能を計測した．$\theta_c = 40$ で正解率は最大で約 59%であった．

CRF による復元は，前後の文脈の範囲を変えて行い，前後 2 単語の場合に最も高い性能を得た．このときの，削除しきい値 θ_c に対する正解率を図 9 に示す．θ_c が大きくなるほど，高い正解率が得られることがわかる．

最後に，削除のしきい値を特定の値に定めて，各復元手法を行い，手法 4 の多数決の性能を調査した．正解率の最大値は 70%であり，CRF の正解率の最大値よりも悪化したため，多数決ではその他の分類器の寄与が小さかったと考えられる．

4.3 自動簡約処理の平均正解精度

前節までで，内容語の自動削除と付属語の自動復元の各々の性能を調べた．本節では，それらの処理を統合した結果として，人手の簡約結果（正解）とどの程度一致しているかを評価する．評価の指標として単語正解精度を用いる．

$$\text{単語正解精度} = \frac{N - (S + D + I)}{N}$$

ここで N は正解における単語の総数，S は置換誤り単語の総数，D は削除誤り単語の総数，I は挿入誤り単語の総数である．前節同様，交差確認法にて平均の値を求める．

簡約処理のパラメータは，内容語の削除に関しては θ_c =40, 70, 90%，付属語の復元の手法 1 に関しては θ_f =50%を採用した．CRF の文脈は前後 2 単語である．

結果を表 2 に示す．しきい値 θ_c がどの値でも，多数決よりも CRF の方が正解精度が高い結果となった．θ_c =90%の場合に正解精度が最も高く（78%），このときの圧縮率は 64%であった．

4.4 講演音声認識結果への適用

前節までに述べた簡約処理を音声認識結果に適用した．使用した講演音声は京都大学 iPS 細胞研究所 (CiRA) の 2010 年シンポジウムにおけるものである．講演の音声認識には，本研究室で開発・公開しているオンライン字幕作成システム [14][*3] を用いた．音響モデルは日本語話し言葉コーパス (CSJ) の講演音声 257 時間で学習した DNN-HMM，言語モデルには CSJ 学会・模擬講演の書き起こし（合計 7.7M 単語）と CiRA の Web サイトから収集したテキスト（合計 53K 単語）から学習した単語 3-gram モデルを用いた．簡約処理のパラメータは，前節同様 θ_c =40, 70, 90%，θ_f =50%を採用した．

評価の指標として圧縮率を次式のように定義する．

$$\text{圧縮率} = \frac{\text{簡約処理後の単語数}}{\text{原文の単語数}}$$

各 θ_c におけるそれぞれの付属語復元手法の圧縮率を表 3 に示す．θ_c=90 では，CRF により 69%に圧縮することができた．それぞれの復元方法で簡約処理した結果の例を図 10 に示す．手法 1〜3 では復元されるべきものが復元されなかったり，復元すべきでないものが復元されてしまう付属語が見られるが，手法 4 の多数決により，バランスのよい簡約結果が得られることがわかる．

[*3] http://caption.ist.i.kyoto-u.ac.jp/

表 2　各 θ_c における各復元手法ごとの平均正解精度と平均圧縮率

		ルールベース	N-gram	CRF	多数決
$\theta_c = 40$	精度 (%)	64.03	59.45	**71.75**	70.88
	圧縮率 (%)	66.77	56.48	60.94	58.42
$\theta_c = 70$	精度 (%)	60.93	53.62	**77.42**	75.49
	圧縮率 (%)	75.72	73.09	63.81	64.19
$\theta_c = 90$	精度 (%)	59.81	51.24	**77.94**	75.83
	圧縮率 (%)	77.35	76.21	64.07	64.81

表 3　音声認識結果を簡約処理したときの各 θ_c における圧縮率

		原文	ルールベース	N-gram	CRF	多数決
$\theta_c = 40$	単語数	5298	3404	2691	3301	3096
	圧縮率 (%)	100.00	64.25	50.79	62.31	58.44
$\theta_c = 70$	単語数	5298	3940	3574	3664	3654
	圧縮率 (%)	100.00	74.37	67.46	69.16	68.97
$\theta_c = 90$	単語数	5298	4005	3686	3680	3716
	圧縮率 (%)	100.00	75.59	69.57	69.46	70.14

5. おわりに

本稿では，聴覚障がい者への情報保障を目的として，音声認識を用いて講演字幕をリアルタイムに作成するための自動簡約処理手法を提案した．具体的には，認識された単語を講演内容を理解するにあたって必要な語（内容語）とそうでない語（付属語）に分け，後者を削除し前者のみを残して字幕として提示する．ただし，これでは必ずしも自然な文にならないことから，例外処理として内容語で削除するものをアノテーション頻度の比率に基づいて決定し，逆に付属語で復元するものをアノテーション頻度の比率，N-gram による言語尤度比較，機械学習を用いる方法により決定する．交差確認法により，削除，復元の適切なパラメータを決定し，講演の音声認識結果に対して簡約処理を行った結果，正解率 78%・圧縮率 64%で文を圧縮することができた．

謝辞

本研究の一部は科学研究費補助金 16H02847 による．

参考文献

[1] 吉川あゆみ，太田晴康，広田典子，白澤麻弓：大学ノートテイク入門—聴覚障害学生をサポートする，人間社 (2001).
[2] 斉藤佐和（監修），白澤麻弓，徳田克己：聴覚障害学生サポートガイドブック—ともに学ぶための講義保障支援の進め方，日本医療企画 (2002).
[3] 勝丸徳浩，河原達也，秋田祐哉，　森信介，　山田篤：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，WIT-109-260 (2009).
[4] 桑原暢弘，秋田祐哉，河原達也：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，日本音響学会春季研究発表会講演論文集，2-4-5 (2014).
[5] 藤井康寿，山本一公，中川聖一：文レベル情報と複数仮説を用いた音声認識結果の自動整形，日本音響学会春季研究発表会講演論文集，2-7-9 (2012).

原文

パーキンソン 病 で は ドーパミン 神経 が 減る
だ から の 中 の ドーパミン が 減る

手法 1: ルールベースによる復元

パーキンソン 病 で ドーパミン 神経 が 減る
から の 中 の ドーパミン が 減る

手法 2: N-gram による復元

パーキンソン 病 ドーパミン 神経 が 減る
中 ドーパミン 減る

手法 3: CRF による復元

パーキンソン 病 で は ドーパミン 神経 が 減る
中 の ドーパミン が 減る

手法 4: 多数決による復元

パーキンソン 病 で ドーパミン 神経 が 減る
中 の ドーパミン が 減る

図 10　講演認識結果の簡約例

[6] Neubig, G., Akita, Y., Mori, S. and Kawahara, T.: A Monotonic Statistical Machine Translation Approach to Speaking Style Transformation, *Computer Speech and Language*, Vol. 26, No. 5, pp. 349–370 (2012).
[7] 河原達也：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 (2012).
[8] Knight, K. and Marcu, D.: Statistics-based Summarization Step one: Sentence compression, *Proc. AAAI/IAAI*, pp. 703–710 (2000).
[9] Jing, H.: Sentence Reduction for Automatic Text Summarization, *Proc. Applied Natural Language Processing (ANLC)*, pp. 310–315 (2000).
[10] Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L. and Vinyals, O.: Sentence Compression by Deletion with LSTMs, *Proc. EMNLP*, pp. 360–368 (2015).
[11] Ohno, T., Matsubara, S., Kashioka, H. and Inagaki, Y.: Simultaneous Summarization of Japanese Spoken Monologue for Real-time Captioning, *Proc. Int'l Conf. Natural Language Processing and Knowledge Engineering (NLP-KE)*, pp. 373–380 (2007).
[12] Furui, S., Kikuchi, T., Shinnaka, Y. and Hori, C.: Speech-to-text and speech-to-speech summarization of spontaneous speech, *IEEE Trans. Speech and Audio Process.*, Vol. 12, No. 4, pp. 401–408 (2004).
[13] 佐藤賢昭，大庭隆伸，政瀧浩和，青野裕司：音声認識結果に対する認識誤りを考慮した教師なし文圧縮，電子情報通信学会技術研究報告，SP2015-84 (2015).
[14] 秋田祐哉，三村正人，河原達也：音声認識を用いた講義・講演の字幕作成・編集システム，情報処理学会研究報告，SLP-108-2 (2015).