講演映像コンテンツ流通のための読みやすい字幕テキスト生成に関する研究

松　原　茂　樹　　名古屋大学情報基盤センター准教授

1 はじめに

聴覚障害者や高齢者，外国人等による講演の理解を支援するための環境として，リアルタイム字幕生成システムの開発が望まれている．１文が長い傾向にある講演文を，そのまま字幕として生成し表示しても必ずしも読みやすくなく，テキストを適切に編集して提示することが必要となる．

そこで本研究では，映像として配信される講演に対して、読みやすい字幕を生成するための要素技術として，まず、講演テキストの改行手法を開発した．本手法では，解説や講演などの字幕情報の提供手段として，字幕のみが複数行表示される画面を想定し，形態素，係り受け，節境界等の出現パタンに基づく改行挿入ルールの適用により，適切な位置で改行された字幕を生成することを試みた．本研究では，形態素情報，文節境界情報，係り受け情報，節境界情報が付与されたデータ[Ohno 2006] に対して，改行位置を人手で付与し，講演文に対する改行データを作成した．次に，改行データの詳細な分析に基づき，改行位置を決定するための改行ルールを定めた．講演の書き起こしテキストを用いて改行挿入実験を実施した結果，人手で改行位置を付与した正解データに対して 86.8%の挿入精度を達成し，本手法の利用可能性を確認した．

本研究では続いて、読みやすい字幕テキスト生成のための技術として、講演テキストにおける読点の挿入手法を実現した．読点挿入に関する研究としてこれまでに，ルールベースで挿入する手法が開発されているものの[鈴木 1995，林 1994]，ルールの数は必ずしも十分でないという問題がある。読点にはいくつかの用法が存在し，用法ごとに挿入位置が異なる．そのために，まず本研究では，読点の挿入位置，用法に関する文献[本多 1982，犬飼 2002，小学館 2007]を調査し，読点の用法を 9 種類に分類した．機械学習に用いるための有効な素性について検討するため，新聞記事中の読点の挿入位置を，分類した読点の用法に従って分析した．本手法では，形態素解析，文節まとめ上げ，節境界解析，係り受け解析が与えられた文を入力とし，入力文中の各文節境界に対して，その位置に読点を挿入するか否かを同定する．京都コーパス[河原 2002] を用いた読点の挿入実験を行った．4,659 文に対して読点挿入を実行した結果，再現率で 69.13%，適合率で 84.13%を達成した．比較のための設定したベースライン手法と比べ性能が向上しており，本手法の有効性を確認した．

本稿の構成は以下の通りである。続く 2 章では、講演テキストへの改行挿入について述べる。3 章ではテキストへの読点挿入手法を説明する。最後に 4 章で、本研究の成果をまとめる。

2 講演テキストへの改行挿入

本研究では，解説や講演などの映像に字幕が提示される環境の実現を目指している。図 1 の講演テキストに対して改行が適切に挿入された字幕例を図 2 に示す．テレビ番組のクローズドキャプションとは異なり，テキストが行単位で入れ替わり，スクロールされながら表示される字幕の生成を目指す。

アメリカでは消費者教育を受ける権利つまり
合理的で賢い意思表示をして市場に参加でき
るための教育を誰もが受けられることこの消
費者教育を受ける権利が消費者の権利の一つ
として付け加えられたのが千九百七十五年で
す

図 1: 字幕テキスト

字幕生成のための改行挿入に関する従来研究として，門馬らは， テレビ番組におけるクローズドキャプションの改行位置を、形態素列のパタンにより改行位置を決定する手法を提案している[門馬 2001]．しかし，日本のテレビ番組におけるクローズドキャプションでは，１画面２行の字幕を一度に切り替える表示方式が

標準となっており，複数行の字幕の段階的な表示環境とは，挿入すべき改行の位置は異なる．

また，西光らは，多段階にチャンキングした言語的なまとまりに基づいて字幕を生成する手法を提案している[西光 2006]．この手法では，文の主題や述語，格要素などに対応する「構成要素」と複数の構成要素からなる「フレーズ」にチャンキングする．しかし，人手による改行位置と「構成要素」や「フレーズ」との関係性は検証されていない．

アメリカでは
消費者教育を受ける権利
つまり合理的で賢い意思表示をして
市場に参加できるための教育を
誰もが受けられること
この消費者教育を受ける権利が
消費者の権利の一つとして
付け加えられたのが千九百七十五年です

図 2：適切に改行された字幕テキスト

2-1 講演テキストにおける改行の挿入位置

複数行の字幕が表示される画面では特に，改行を考慮することなく字幕を生成し表示すると読みにくいテキストとなるため，適切な位置で改行する必要がある．本研究では，字幕生成における適切な改行挿入位置として，以下の考え方を設けた．

- 各行が意味的なまとまりを形成するように改行する．ただし，連体節については意味的にまとまっているものの，そこで改行すると，行末を文末と誤解される恐れがあるため，改行の対象外とする．
- ディスプレイの大きさと文字の可読性を考慮した行の最長文字数を設定し，それ以下となるように改行する．行によって長さが大きく異なると閲覧性が低下すると考えられるため，各行の文字数がおおよそ均等となるように改行する．

なお，本研究では，改行の挿入位置は文節境界に限定する．

2-2 講演テキストにおける改行位置の検出

提案手法では，形態素解析，文節まとめ上げ，節境界解析，係り受け解析が施された文を入力とし，改行位置（以下，改行点）が付与された文を出力とする．一行が最長文字数以下となるように，意味的なまとまりの切れ目を手がかりに改行点を特定する．本手法における改行点の検出プロセスは以下の通りである．

1. 改行点の候補選定改行点検出ルールを適用することにより，一行が最長文字数を超えない程度の数の改行点を挿入し，それを改行点候補とする．
2. 改行点の決定一行が最長文字数を超えない範囲で，可能であれば改行点候補を削除し，最終的に残った候補を出力する改行点として決定する．

なお，本研究では，2.1 節で示した考え方に基づいて，独話文の改行点を検出するルール（以下，改行ルール）を作成した．改行ルールは，入力文における意味的なまとまりを捉えるという観点から，節境界，及び，係り受けに着目して定めた．

以下ではまず，作成した改行ルールについて説明し，次に最終的な改行点決定法について述べる．

表 1：最も改行点になりやすい節境界の種類

並列節	トカ，ガ，シ，デ，ケレドモ，タリ
条件節	カギリ，バ，タラ，ケッカ，トコロ
時間節	トキニ，アトニ，イマ，アト，トキノ，その他
理由節	ノデ，カラ
その他	体言止，間投句，連用節その他，タメニハ節 タメ節，ナガラ節，ナド節，テハ節，ヨウ節

表 2：改行点になりやすい節境界の種類

間接疑問節，補足節，譲歩節テモ，連用節

（１）節境界に基づく改行

節境界とは，隣接する節間の境界である．一般に，節は，意味的なまとまりを形成するため，節境界は改行点の有力な候補となる．ただし，あらゆる種類の節境界が必ずしも改行点になりやすいわけではない．我々の調査では，最も改行点になりやすい節境界の種類 2 は表 1 に示す通りである．すなわち，

【改行ルール 1】 表 1 に示した種類の節境界に改行点を挿入する。

を設ける．改行データに出現した節境界は，全部で 59 種類存在しており，改行点になる節境界は 49.1%を占めている．ただし，ルール 1 だけで，「一行が最長文字数を超えない」という条件を満たさない場合には，上述の節境界の次に改行点になりやすい節境界での改行を実施する．すなわち，

【改行ルール 2】表 2 に示した種類の節境界に改行点を挿入する。

を設ける．

（２）係り受け関係に基づく改行

係り受け関係とは，係り受け文節と受け文節の間の関係であり，ある種の意味的なまとまりを形成する．そのため，係り受け関係の直後には，節境界に準じて改行が挿入されやすい．同時に，係り文節と受け文節の間に改行が挿入される場合，文節の係り先の解釈を読み手が誤る可能性があり，それを考慮した改行点を選ぶ必要がある．

係り受け関係のうち，最も改行が挿入されやすい場合は，連体節の係り先の直後である．加えて，連体節で改行を行うと，行末を文末と誤解される恐れがあるため，

【改行ルール 3】連体節の最終文節の受け文節の直後に改行点を挿入する

を設ける．また，文の冒頭に節境界「主題ハ」が存在し，かつ，述語が複数存在する場合には，「主題ハ」の文節は文末に係ることが多い．その場合，複数の述語の間で行を分割してしまうと，「主題ハ」の文節の係り先が前者の述語であると，読み手が誤って解釈する恐れがある．これを避けるために，節境界「主題ハ」の直後を改行の候補とする．すなわち，

【改行ルール 4】文の冒頭に，文末に係る節境界「主題ハ」が存在し，かつ，述語が複数存在する場合には，「主題ハ」の直後に改行点を挿入する．

を設ける．その他に，係り文節と受け文節が離れて位置する場合などにおけるルールとして，

【改行ルール 5】係り文節と受け文節の間に改行が挿入される場合，係り文節の直後に改行点を挿入する．

を設ける．

以上の改行ルールを手がかりに改行点を検出する．なお，番号は，ルールの優先性を示しており，例えば，改行ルール 1 は最も強力であり，該当の位置には無条件に改行が挿入される．

（３）改行点の決定

改行点の候補選定では一行の文字数を最長文字数に近づけることはあまり考慮されていないため，改行点候補をそのまま改行点として確定すると短い行が連続して生じる場合がある．改行点の決定処理では，改行点候補をそのまま改行点として確定した場合に生じる連続した短い行を一行に連結できるか否かを判定し，結合できる場合には，連結する行間の改行点候補を削除する．

まず，連続する二行を結合した文字列が最長文字数を超える場合は，これら二行は結合できないと判定する．次に，たとえ，結合した文字列が最長文字数以内であっても，改行点候補の位置が意味的に強い切れ目であれば，必ず改行される必要があり，そこでは結合できないと判定する．さらに，改行点候補の位置が意味的に強い切れ目でない場合でも，一行が意味的なまとまりとなることが望ましいため，「連続する二行の最終文節が共に同じ文節に係る」，もしくは，「結合した文字列内で係り受け構造が閉じている」場合のみ，結合できると判定する．

なお，以下の条件を全て満たす場合，改行点候補の位置は意味的に強い切れ目でないとする．

- その箇所が節境界でない．
- その箇所の直前の形態素が，「係助詞モ」もしくは「名詞」，「副詞」，「場合(名詞-副詞可能)」でない．

2-3 実験

本手法の利用可能性を確認するため，実際の講演データに対して改行挿入実験を実施した．

（１）実験の概要

実験データとして，３つの日本語講演の書き起こしテキスト（219 文，2121 文節）を使用した．データには，節境界情報，形態素情報，係り受け情報を人手で付与している．実験データに対して，本手法による改行挿入実験，また，文節単位ごとに文字列を連結していき，文字数が最長文字数に最も近づいた時点で改行を挿入するという手法をベースラインとして改行挿入実験を実施し，結果を比較した．

改行挿入結果に対する評価は，正解の改行位置に対する精度（挿入した改行のうち正解と一致する割合），及び，再現率（正解の改行位置に改行が挿入された割合）を測定することにより実施した．正解データは，二人の作業者が共同で作成した．すなわち，まず二人が独立に正解データを作成し，それぞれのデータをもとにした協議を経て，正解データを作り上げた．生成する字幕として容認可能な改行位置が正解データ以外に存在することを考慮し,正解データの作成前に二人の作業者が別々に作成した２つの改行データを使用し，それらとの一致に基づく精度も併せて測定した．

（２）実験結果

実験結果を表 3 に示す．精度で 78.4%，再現率で 78.7%と，ベースラインと比べて高い性能を示した．また，正解とは異なる２つの改行データを対象に評価したところ，86.8%の精度を達成した．以上より，本手法の利用可能性を確認した．挿入誤りを分析した結果，

- 隣接する文節「考えてみたいと」と「思います」の間のように，改行点挿入の可能性が，文法とは別に，語彙の種類によって決まる場合が存在する．
- 改行決定プロセスでは，削除する改行点候補の順序により，最終的な改行点が異なってくる．

などの原因が明らかになった．

表 3: 実験結果

	本手法	ベースライン
再現率	78.7%（406/515）	30.1%（155/515）
精度	78.4%（406/517）	37.9%（155/409）

3 講演テキストへの読点挿入

3-1 読点位置の分析

本研究では，表 4 の分類に従い，読点の用法に注目した読点の挿入手法の開発を目指した。適切な読点位置とは，いくつかの要因のバランスのもとに定まると考えられるため，本研究では統計的アプローチを採用する．機械学習のための有効な素性について検討するため，事前分析を与えた．

表 4 読点の用法の分類

節間に打たれる読点
係り受け関係を明確にする読点
難読・誤読を避ける読点
主題を示す読点
先頭の接続詞・副詞の後に打たれる読点
並列する単語・句の間に打たれる読点
時間を表わす副詞の後に打たれる読点
直前の語句を強調するための読点

分析には，京都テキストコーパス 4.0（以下，京都コーパス）[河原 2002] の 1 月 1 日，及び，3 日から 11 日までの全記事を用いた.使用した分析データの規模を表 5 に示す.コーパス中のテキストには,形態素，文節境界，係り受け構造の構文的情報が，人手により付与されている．また，節境界解析ツール CBAP[丸山 2004] を用いて節境界情報を自動で付与した．

表 5 分析データの規模

文数	11,821
文節数	117,501
文字数	503,970
読点数	16,595
平均文長	42.63

読点のうち， 文節境界以外（すなわち，文節内）に挿入されているものは全体の 1.43%(238/16,595）に

過ぎなかった．そこで，文節境界に挿入されている読点のみを分析の対象とした．文節境界 105,680 箇所に対する読点挿入率は 15.48%（16,357/105,680）である．分析では，それぞれの読点の用法ごとに，形態素や係り受け構造，節境界，読点によって挟まれた文節列の文字数の情報に注目し，それらと読点位置との関係について調査した．なお，「直前の語句を強調するための読点」については，執筆者の意図に依存するものであるため，本研究では対象としない．

（1）節間に打たれる読点

読点を節と節の間に打つことにより文の構造が分かりやすくなる．このことから，節の境界は読点位置として有力であると考えられる．例えば以下の文

- 国連による対イラク制裁解除に向け、関係の深い仏に一層の協力を求めるのが狙いとみられる。

では，文節「向け」の直後に存在する節境界「連用節」に読点が挿入されている．分析データでは，文末を除く節境界 29,278 箇所のうち 8,805 箇所に読点が挿入されており，節境界に対する挿入率は 30.01%であった．文節境界に対する挿入率よりも高いことから，節境界には読点が挿入されやすいといえる．

分析データに出現した 114 種類の節境界について，種類ごとに読点挿入率を調査した．節境界「連用節」や「並列節ガ」，「並列節デ」の読点挿入率は 80%を越えているのに対して，「連体節」や「引用節」には 5%以下しか読点は挿入されていなかった．これらは，節境界の種類によって読点の挿入されやすさが異なることを示している．

（2）係り受け関係を明確にする読点

読点には係り受け関係を明確にする働きがある．実際，分析データを調査したところ，係り受け関係にある隣接文節間 66,984 箇所に対して，読点が挿入されたのは 2,302 箇所で，挿入率は 3.44%に過ぎなかった．一方，係り受け関係にない隣接文節間への挿入率は 36.32%であった．また，係り受け構造と読点との関係，すなわち，読点によって挟まれた文節列内で係り受けが閉じているかどうかを調べた．ここで，係り受けが閉じている文節列とは，文節列外の文節に係る文節が，文節列末の文節以外に存在しない文節列のことをいう．読点に挟まれた文節列 16,357 個のうち，12,496 個(76.40%) で係り受けが閉じていた．この結果も，係り受け距離が遠くなる文節の直後には読点が挿入されやすい傾向を反映している．

ある文節の係り先が節末の文節よりも遠い場合，その文節の係り先を明確にするため読点が挿入されやすいと考えられる．例えば，以下の例では，文節「響き」の直後に節境界「連用節」が存在しており，文節「中心に」が節境界の直前の文節「響き」よりも遠くの文節「占拠」に係っている．そのため，文節「中心に」の直後に読点が挿入されている．

- モガディシオからの報道によると、市内のベルムダ地区を中心に、ロケット弾の爆発音が響き、通りを武装兵士が占拠。

隣接する文節が所属する節の節末文節よりも遠くの文節に係る文節の境界に対する読点挿入率を測定した．挿入率は 54.24%（7,478/13,786）であり，文節境界全体と比べて読点が挿入されやすい．

（3）難読・誤読を避ける読点

漢字やカタカナが続けて出現すると，読み手が誤読をしたり，読みづらさを感じたりすることがある．それを避けるために，このような文節境界には読点が挿入される．以下の例では「営業マン」と「マイケル・スタメンソン氏とともに」の間の読点が，誤読・難読を避けるために挿入されている．

- 出納責任者ロバート・Ｌ・シトロン氏は、アドバイザーでもあったメリル・リンチ証券の営業マン、マイケル・スタメンソン氏とともに、米証券取引委員会から事情聴取を受けている模様。

文節にまたがって漢字が出現するような文節境界 2,409 箇所のうち 90.83%(2,188/2,409)に，また，カタカナの場合は 97.69%(211/216) に読点が挿入されていた．文節にまたがって漢字やカタカナが連続する場合，そのほとんどの文節境界に読点が打たれる傾向にある．

（4）主題を示す読点

文の主題を示すような文節の直後には，主題を明確にする目的で読点が打たれやすいと考えられる．例えば，

- その最大の理由は、香港町が低空飛行を続けるカナダ・トロント経済を活性化している

という文では，主題を示す文節「理由は」を明確にするために，その直後に読点が挿入されている．そこで，節境界「主題ハ」に注目して分析を行った．節境界「主題ハ」への読点挿入率は 16.94%（1,446/8,536）であり，文節境界に対する読点挿入率との差は小さい．これは，単純に主題を表わす文節の直後に読点を挿入すると，読点の数が多くなり文が読みにくくなることから，読点が挿入されないことも多いためであると考えられる．しかし，節境界「主題ハ」に挿入されている読点は，読点全体の 8.84%（1,446/16,357）を占め

ている．

隣接する文節に係らない文節の直後に存在する節境界「主題ハ」への読点挿入率は 20.71%であり，「主題ハ」全体に対する読点挿入率よりも高い．隣接しない文節に係る文節の直後に存在する「主題ハ」には読点が挿入されやすいといえる．また，以下の例の「報道では」のように，節境界「主題ハ」の直前の文字列が「では」であった場合，読点挿入率は 35.82% (254/709) であり，「主題ハ」への読点挿入率よりも高い．

● グロズヌイからの報道では、ロシア軍は激しい空爆と砲撃を加えた後、装甲軍部隊が大統領官邸付近に進出。

（5）先頭の接続詞・副詞の後に打たれる読点

以下の例の「しかし」のように，文頭に出現する接続詞や副詞の直後には前置きの語を区切るという目的で読点が挿入されることが多い．

● しかし、旧民社党は大半の議員が新進党に参加し、さきがけとの連携も流動的で連携相手は不確定だ。

分析データ中で，最終形態素が「接続詞」である文頭の文節の直後には，71.65%(498/695)の確率で読点が挿入されていた．また，最終形態素が「副詞」である場合では，挿入率は 30.97% (140/452) であった．文節境界に対する読点挿入率よりも高い挿入率であることから，文頭に出現する前置きの語の直後には読点が挿入されやすいといえる．

（6）並列する単語・句の間に打たれる読点

読点には，対等の関係で並列された同じ種類の語や句を区切るという働きがある．以下に例を示す．

● むしろ地球規模の環境、人口、食糧など広範に国連の果たさなければならない役割は大きい。

この例では，「環境」「人口」「食糧」と並列された名詞を区切るためにその間に読点が挿入されている．最終形態素が名詞である文節が連続する場合，その文節境界への読点挿入率は 59.39% (3,330/5,607) であった．また，語が並列される以外に句が並列される場合がある．以下の例

● メニューは前夜、首相が何を食べたかを調べて同じ献立を避けたり、和食と洋食のバランスを考えたりして決める。

では，「同じ献立を避けたり」という動詞句と「和食と洋食のバランスを考えたり」という動詞句が並列されているため，動詞句の並列を明確にするために，文節「避けたり」の直後に読点が打たれている．分析データ中で，文末の述語に係る文節 B（動詞）が存在し，文節 B の後方に，文末の述語に係る文節（動詞）が存在する場合の文節 B への読点挿入率を調査した．例えば，上記の例文では，文節「避けたり」は文末「決める」に係り，かつ，その後方に同じく「決める」に係る文節「考えたりして」が存在するため，調査対象に該当する．

調査の結果，読点挿入率は 79.89%(751/940) であり，文節に対する読点挿入率と比較して大幅に高い値であった．すなわち，節が並列された場合には読点が打たれやすいといえる．

（7）読点によって挟まれた文節列の文字数

読点には表 1 に示したように様々な用法があるため，むやみに多く打たれる訳ではない．そのため，読点によって挟まれた文節列の文字数が短くなりすぎることはない．文字数が 1 文字である文節列の出現頻度は 50 回以下である．文節列の文字数が 4 文字，5 文字の場合は頻度が約 400 回であり，6 文字から 21 文字までは頻度が 400 回を超えていた．22 文字以降は頻度が下がっていくという結果が得られた．この分析結果に従って，読点によって挟まれた文節列の文字数を 4 分類し，素性として用いた．

3-2 統計的な読点挿入手法

本手法では，形態素解析，文節まとめ上げ，節境界解析，係り受け解析が与えられた文を入力とし，入力文中の各文節境界に対して，その位置が読点位置であるか否かを同定する．3-1 の分析において，読点の 98.57%が文節境界に挿入されていたことから，本手法では文節境界のみを読点位置の候補とした．入力文に対する適切な読点位置を同定するために，一文において考えうる読点位置の全ての組み合わせの中から，最適な組み合わせを、確率モデルを用いて決定する．

表 6 テストデータの規模

文数	4,659
文節数	46,511
文字数	198,899
読点数（文節境界）	6,549
平均文長	42.69

3-3 実験

本手法の有効性を評価するため，日本語テキストデータを用いて読点の挿入実験を実施した．

（1）実験の概要

実験には京都コーパス[河原 2002] に収録されている日本語テキストデータを用いた．テストデータには 1 月 14 日から 17 日の全記事を，学習データには分析データと同一のテキストを使用した．テストデータの規模を表 6 に示す．なお，実験のための最大エントロピー法のツールとしては，文献[Zhang 2008] のものを利用した．オプションに関しては，学習アルゴリズムにおける繰り返し回数を 2,000 に設定し，それ以外はデフォルトのまま使用した．評価は，正解の読点位置に対する再現率と適合率，及び，それらの調和平均である F 値により行った．

比較のために，節や係り受けの情報などを考慮せず，形態素情報のみを用いて読点を挿入する手法をベースラインとして設定した．ベースライン手法では素性に，文節の主辞（品詞，活用形），語形（品詞，表層文字）と隣接文節の第一形態素（品詞，表層文字）を用いた．

（2）実験結果

提案手法ならびにベースラインの再現率，適合率及び F 値を表 7 に示す．提案手法は，再現率で 69.13%，適合率で 84.13%を達成した．F 値の比較において，提案手法はベースラインと比較して高い性能を示しており，提案手法の有効性を確認した．

表 7 実験結果

	再現率	適合率	F 値
提案手法	69.13%（4,527/6,549）	84.13%（4,527/5,381）	75.90
ベースライン	51.38%（3,365/6,549）	70.90%（3,365/4,746）	59.58

提案手法による読点挿入結果のテキストの例を図 3 に示す．節間や並列する名詞の間，文頭の接続詞の直後や主題を示す「は」の直後などに正しく読点が挿入されていることがかる．すべての読点位置が一致した文数の割合を示す文一致率は 55.81%（2,600/4,659）であり，半数以上の文で，一文中の全ての読点位置に正しく読点を挿入できた．

- しかし、激しい雪の影響で復旧作業は進まず、直江津発上野行きの特急「あさま３８号」が妙高高原駅で停車したままとなったのをはじめ、計七本が長野、新潟両県内の駅で立ち往生。
- 九、十の両日行われたジュネーブ会談の成果として、当事者などを含めた拡大会談の開催が決まったが、アラタス外相は「この拡大会談は、国連事務総長のリーダーシップで行われるものではなく、東ティモール人自身の前向きな姿勢で可能となるものである」と強調し、五月十九日にニューヨークで開く次の会談にホルタ氏の参加を促した。
- 村山富市首相は十六日昼、首相公邸で、さきがけ代表の武村正義蔵相と会談、今後の社会、さきがけ、両党による連携強化や社会党の新民主連合による新会派結成問題について協議する。

図 3 提案手法による読点挿入結果のテキストの例

提案手法による読点挿入結果のテキストとベースラインによる読点挿入結果のテキストを図 4 に示す．ベースライン手法では，文節「浮かんでいるが」の直後や「決まらないため」の直後に読点が挿入されていない，あるいは、「名乗る」と「副司令官」の間のように不自然な位置に読点が挿入されている．一方，提案手法ではそのような位置に正しく読点が挿入できている．

（提案手法）

候補者として石原信雄内閣官房副長官や岩國哲人・島根県出雲市長、鳩山邦夫前労相、作家の堺屋太一氏らの名前が浮かんでいるが、前提となる政党の枠組みが決まらないため、調整は難航。

（ベースライン）

候補者として石原信雄内閣官房副長官や岩國哲人・島根県出雲市長鳩山邦夫前労相、作家の堺屋太一氏らの名前が浮かんでいるが前提となる政党の枠組みが決まらないため調整は難航。

図 4 提案手法とベースラインによる読点挿入結果の比較

正解の読点位置のうち，ベースライン手法では読点が挿入されず，提案手法のみ挿入した箇所は 1,603 箇所であった．一方，ベースライン手法によってのみ正解の読点位置に読点を挿入できた箇所は 441 箇所であった．ベースラインで挿入できていなかった種類の読点が提案手法で挿入できるようになったというわけではなく，それぞれの読点の用法に関する素性を用いることによって，読点挿入の性能が全体的に向上したといえる．

3-4 考察

（1）読点挿入誤りの分析

正解の読点位置のうち，読点が挿入されなかった箇所は 2,022 箇所であった．正解の読点位置に挿入されなかった箇所のうち，862 箇所は節境界であり，節境界「主題ハ」がその 53.36% (460/862) を占めていた．節境界「主題ハ」は出現数が多く，読点が挿入される数も多くなる．しかし，読点挿入率自体はそれほど高くない．本手法では，節境界「主題ハ」に関する素性を 4 種類導入しているが，それらが必ずしも有効に働いていなかったといえる．表 8 に節境界「主題ハ」に対する読点挿入の再現率及び適合率を示す．実際，テストデータ中で，「主題ハ」に挿入されている読点は 601 箇所存在したが，そのうち正しく挿入できた箇所は 141 箇所であった．「主題ハ」への読点挿入をより正しく行うための素性の検討が今後の課題となる．読点が挿入できなかった節境界のうち，「主題ハ」に次いで多かったのは節境界「テ節」であり，108 箇所であった．

表 8 節境界「主題ハ」に対する読点挿入結果

再現率	適合率	F 値
23.46%(141/601)	59.49%(141/237)	33.65

節境界以外では，連続する名詞間に読点が挿入されなかった箇所が 130 箇所存在した．以下にそのような読点挿入結果の例を示す．

- （正解）ボウルに豚の背脂、ニンニク、ショウガ、ネギのみじん切りを入れ、彩りの赤ピーマンも加えます。
- （提案手法）ボウルに豚の背脂ニンニク、ショウガ、ネギのみじん切りを入れ、彩りの赤ピーマンも加えます。

上記の例で，正解データでは，文節「背脂」と「ニンニク」の間に読点が挿入されている．提案手法では，文節「背脂」の直後に読点が挿入された場合，読点間の距離が短くなることから，読点によって挟まれた文節列の文字数に関する素性が悪影響を及ぼした可能性がある．一方，文節「ニンニク」と「ショウガ」の間には正しく読点が挿入されているが，これは名詞が連続していることの他に，カタカナが文節にまたがって出現しているため，読点が正しく挿入されたと考えられる．

（2）不自然な読点挿入

読点位置の中には，正解の読点位置とは異なっても許容できる読点位置が存在する．しかし，明らかに不自然な位置に読点が挿入された場合，文の意味が変わったり，読み手が係り受け構造を誤って認識したりするなど，その読点挿入誤りが与える影響は大きい．そこで，提案手法によって明らかに不自然な位置に挿入されている読点位置を調査した．1 月 14 日の記事中の 217 文（2,349 文節）に対する読点挿入結果のうち，提案手法が正解と異なる位置に挿入した読点 47 箇所とした．読点が不自然であるか否かの判定は，3 名の作業者による協議のもと決定した．

調査の結果，明らかに不自然な読点挿入位置と判定したのは，47 箇所のうち 4 箇所であった．以下で，その 4 箇所の読点挿入位置と，不自然と判定した要因について述べる．

- 政党助成を行う主要国の例は、ドイツが年間約百五十億円で、国民一人当たり百八十四円▽フランス、約百五億円同百八十三円▽スウェーデン、約十九億円同二百十七円――などだ。

この文では，文節「フランス」の直後と「スウェーデン」の直後に挿入されている読点を不自然と判断した．この 2 箇所の位置に読点を挿入することによって，「約百五億円同百八十三円▽スウェーデン」が一つのまとまりに見えてしまうためである．

- 同省の特殊法人は計十三、あり、所管省庁別では運輸省に次いで多い。

上記の文では，「十三あり」が一つのまとまりであるにも関わらず，「十三」と「あり」の間に読点が挿入されている．人間はこのように明らかに不自然な位置に読点を挿入することはないと考えられる．

- 日米首脳会談のため、訪米していた村山富市首相は十三日午後二時前、政府専用機で羽田空港に到着した。

上記の文において，文節「ため」は直後の文節「訪米していた」に係る．しかし，「ため」の直後に読点が挿入されることによって，「ため」が遠くの文節に係ると読み手が錯覚する可能性がある．文の係り受け構造を誤って認識する可能性があることから，この読点は不自然であると判定した．

不自然と判断された箇所が 47 箇所のうち 4 箇所であったことから，提案手法は，ある程度自然な位置に読点を挿入できているといえる．

（3）人間による読点挿入の一致率

実験では，正解データとの比較によって読点挿入の結果を評価した．しかし，実験結果の値がどの程度の値を示せば十分なのかは定かではない．そこで，人間による読点挿入作業を行い，その結果を一つの指標とし，提案手法の読点挿入性能を評価した．

上述のデータと同様の 217 文に対して，日本語の文章作成に精通する作業者 1 名が読点挿入を行った．正解データに対する作業者，及び，同様のデータに対する提案手法の再現率，適合率とその F 値を表 9 に示す．人間の作業においても F 値は 79.42 であり，読点挿入作業は人間でも揺れが生じるタスクであることが分かる．提案手法は F 値において，作業者の読点挿入性能の 96.30%（76.48/79.42）を達成している．また，提案手法は精度で 82.78%を示していることから，揺れが生じる読点作業において，適切な読点挿入が行えていることがわかる．

表 9 人間による読点挿入

	再現率	適合率	F 値
作業者	78.30%（249/318）	80.58%（249/309）	79.42
提案手法	71.07%（226/318）	82.78%（226/273）	76.48

4 おわりに

本稿では，聴覚障害者，高齢者，外国人等による独話音声の理解を支援することを目指し，字幕生成における独話文への改行挿入手法を提案した．本手法では，形態素，係り受け，節境界，ポーズ・フィラー等の出現に基づく改行挿入ルールの適用することにより，字幕テキストを読みやすくするための改行を実現する．解説番組を用いた改行挿入実験では，精度で 78.4%，再現率で 78.7%であり，本手法の利用可能性を確認した．人手により改行挿入ルールを作成する方式では，ルールの大規模化，詳細化，体系化の面で限界がある．現在，改行挿入データの作成を進めており，今後，本研究により得られた知見をもとにルール獲得方式を考案することを検討している．また，本研究で対象とした改行の挿入位置は，音声合成における適切なポーズ挿入位置[佐藤 1999]と関連することが予想される．両者の言語的特徴の共通性について明らかにすることも今後の課題である．

また本稿では，日本語テキストにおける読点の挿入手法を提案した．本手法では，読点の用法に注目し，形態素や係り受け，節境界等の情報に基づき，統計的手法によって一文中の適切な読点の挿入位置を同定する．読点位置の検出実験では再現率と適合率の F 値で 75.90 を示しており，本手法の有効性を確認した．実験結果の分析から，特定の用法の読点が挿入できていないことが分かった．今後は，その用法の読点に関する，より有効な素性を発見・利用し，本手法の再現率を向上することが課題となる．また，本研究では対象としなかった「直前の語句を強調するための読点」に関しても，今後検討する必要がある．

【参考文献】

門馬隆雄，沢村英治，福島孝博，丸山一郎，江原暉政，白井克彦：聴覚障害者向け字幕付きテレビ番組の自動制作システム，電子情報通信学会論文誌，Vol.J84-D-II，No.6，pp.888-897 (2001).

佐藤奈穂子，小島裕一，望主雅子，亀田雅之：テキスト音声合成における係り受け解析結果を用いたポーズ挿入処理，自然言語処理，Vol.6，No.2，pp.117-133 (1999).

西光雅弘，高梨克也，河原達也：係り受けとポーズ･フィラーの情報を用いた話し言葉の段階的チャンキング，情報処理学会研究報告，2006-SLP-61，pp.19-23 (2006).

T. Ohno, S. Matsubara, H. Kashioka, N. Kato, Y. Inagaki: A Syntactically Annotated Corpus of Japanese Spoken Monologue, Proc. of 5th LREC, pp.1590-1595 (2006) .
丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝: 日本語節境界検出プログラム CBAP の開発と評価, 自然言語処理, Vol. 11, No. 3, pp.39-68 (2004).
鈴木英二, 島田静雄, 近藤邦雄, 佐藤尚, 日本語文章における句読点自動最適配置, 情報処理学会全国大会講演論文集, Vol.50, No.3, pp.185-186 (1995).
林良彦, 技術文章向けの日本文推敲支援システムの実現と評価, 電子情報通信学会論文誌, Vol.J77-D-II, No.6, pp.1124-1134 (1994).
本多勝一, 日本語の作文技術, 朝日新聞社出版局 (1982).
犬飼隆, 文字・表記探究法, 朝倉書店 (2002).
小学館辞典編集部編, 句読点、記号・符号活用辞典。, 小学館(2007).
河原大輔, 黒橋禎夫, 橋田浩一,「関係」タグ付きコーパスの作成, 言語処理会第 8 回年次大会発表論文集, pp.495-498 (2002).
L. Zhang: Maximum entropy modeling toolkit for python and c++ (2008). http://homepages.inf.ed.ac.uk/s0450736/maxent/toolkit.html [Online; accessed 1-March-2008].

〈発 表 資 料〉

題 名	掲載誌・学会名等	発表年月
Automatic Linefeed Insertion for Improving Readability of Lecture Transcript	Studies in Computer Science	2009 年 7 月
読点の用法的分類に基づく自動読点挿入	情報処理学会研究報告	2010 年 5 月
Construction of Linefeed Insertion Rules for Lecture Transcript and their Evaluation	International Journal of Knowledge and Web Intelligence	2010 年 6 月