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の評価

藤井康寿[†1,†2]　山本一公[†1]　中川聖一[†1]

我々はこれまでに，認識誤りに頑健な整形手法として，コンフュージョンネットワークで表現される認識結果の複数仮説を扱うことができる音声認識結果の整形手法を提案してきた．しかし，複数仮説を利用することで，文節チャンキングや依存構造解析といった文レベル情報を同時に使用することが難しいという問題があった．本稿では，複数仮説を利用しながら文レベル情報を利用することができる新しい文整形アルゴリズムを提案する．客観評価および主観評価によって，文レベル情報と複数仮説を同時に利用することでより良い整形文を得られることがわかった．

Evaluation of Transcription Cleaning Algorithm for ASR Results using Sentence Level Knowledge and Multiple Hypotheses

Yasuhisa Fujii[*], Kazumasa Yamamoto and Seiichi Nakagawa[†1]

In a previous work, we proposed a method of transcription cleaning for improving the readability which dealt with multiple hypotheses represented by a confusion network to be robust to the recognition errors. However, utilizing multiple hypotheses causes an additional challenge that sentence-level knowledge such as chunking and dependency parsing is difficult to be incorporated. In this paper, we propose a novel algorithm which infers clean, readable transcripts from multiple hypotheses represented by a confusion network while integrating sentence-level knowledge. Objective and subjective evaluations showed that using sentence-level knowledge with multiple hypotheses was effective for cleaning of ASR results.

†1 豊橋技術科学大学
Toyohashi University of Technology
†2 学振特別研究員

1. はじめに

音声の書き起こしが利用可能になれば，音声ドキュメントの内容を理解することがより容易になると考えられ，音声ドキュメントの自動認識の研究が活発に行われている．ユーザの理解を助けるだけでなく，本稿で対象とする講義音声においては，要約[1)]，インデキシング[2)]，およびブラウジングシステム[3)]などの後処理にも書き起こしが必要となるため，講義音声を書き起こすための研究が盛んに行われている[4)–6)]．

しかし，講義音声などの自然発話の書き起こしは，フィラー，言い直し，繰り返し，助詞の脱落といった話し言葉特有の現象が頻繁に発生するため，人手による完全な書き起こしであっても人間にとっては読み辛く，理解し難いものとなる．そのため，書き起こしをユーザに提示する前に，認識結果の整形処理を行い，読み易さおよび理解し易さを向上することが望まれる[7)–10)]．

我々は，認識誤りに頑健な整形手法として，コンフュージョンネットワークによって表現される複数仮説を用いた音声認識結果の自動整形手法を提案し，その有効性を示している[11)]．しかし，文献 11) の方法は，複数仮説を利用することで，句読点の挿入などにおいて文整形にとって有用であると考えられる文レベル情報の利用が困難になるという問題があった．

本稿では，文献 11) の手法をさらに改善するために，複数仮説を用いながら文レベル情報を使用するためのアルゴリズムを提案する．コンフュージョンネットワークで表現される複数仮説を用いながら文レベル情報を使用するために，文献 12) において提案されている iterative decoding を利用する．Iterative decoding を用いることで，複数仮説を用いながら文レベル情報を使用することが可能となる．

本稿では，文レベル情報として，文節チャンキングスコア，依存構造解析スコア，および文境界検出結果を使用する．これらの情報を得るために，文境界が未知の話し言葉系列に対して文境界検出と依存構造解析を同時に行うアルゴリズムである Improved-SDA[13)] を使用した．

客観評価と主観評価の結果，文レベル情報と複数仮説を同時に使用することで，文整形が改善することがわかった．

2. Improved-SDA

本稿では，文レベル情報として，文節チャンキングスコア，依存構造解析スコア，および文境界検出結果を使用する．これらの情報を得るためには，文境界が未知の話し言葉系列に

対して文境界検出と依存構造解析を同時に行うアルゴリズムである Improved-SDA[13] を使用した．以下，improved-SDA を構成する各要素について簡潔に述べる．

2.1 文節チャンキング

Improved-SDA は文節列を入力とするため，予め入力単語列に対して文節検出 (文節チャンキング) を行っておく必要がある．本稿では，文節チャンキングのために，文献 13) と同様の構成で CRF ベースの文節チャンカを実装した．話し言葉の系列 $\boldsymbol{s}$ に対して文節チャンキング結果が $\boldsymbol{b}$ となる確率は $P(\boldsymbol{b}|\boldsymbol{s})$ で表現される．

2.2 依存構造解析

一般的な日本語の依存構造 (係り受け) 解析においては，依存構造 $\boldsymbol{d}$ は修飾辞 (modifier)$\boldsymbol{b} = (b_1, b_2, \cdots, b_N)$ と修飾辞に対応する主辞 (head)$h_1, h_2, \cdots, h_N$ の集合で表現される．依存構造解析のタスクは，文節系列 $\boldsymbol{b}$ が与えられたときに，最も適切な依存構造 $\hat{\boldsymbol{d}}$ を発見する問題である．文節列 $\boldsymbol{b}$ に対して依存構造 $\boldsymbol{d}$ が与えられる確率を $P(\boldsymbol{d}|\boldsymbol{b})$ で表す．$P(\boldsymbol{d}|\boldsymbol{b})$ は以下の式で計算される．

$$P(\boldsymbol{d}|\boldsymbol{b}) = \prod_{i=1}^{N} P(b_i \rightarrow h_{b_i}|\boldsymbol{\Phi}(b_i, h_{b_i}, \boldsymbol{b})) \tag{1}$$

ここで，$\boldsymbol{\Phi}(b_i, h_{b_i}, \boldsymbol{b})$ は修飾辞 b_i，主辞 h_{b_i}，文節列 $\boldsymbol{b}$ から得られる言語素性ベクトルである．$P(b_i \rightarrow h_{b_i}|\boldsymbol{\Phi}(b_i, h_{b_i}, \boldsymbol{b}))$ は b_i と h_{b_i} の間に依存関係が存在するスコアであり，以下のように計算される．

$$P(b_i \rightarrow h_{b_i}|\boldsymbol{\Phi}(b_i, h_{b_i}, \boldsymbol{b})) = \frac{\exp(\sum_g^K \boldsymbol{\mu_g} \sigma(\boldsymbol{w_g} \cdot \boldsymbol{\Phi}(b_i, h_{b_i}, \boldsymbol{b})))}{\sum_{h \in C_{b_i}} \exp(\sum_g^K \boldsymbol{\mu_g} \sigma(\boldsymbol{w_g} \cdot \boldsymbol{\Phi}(b_i, h, \boldsymbol{b})))}, \tag{2}$$

ここで，K はゲート関数の数，μ_g はゲート g の重み，w_g はゲート g に対応する重みベクトルを表す．$\sigma(x)$ はゲート関数で，本稿ではシグモイド関数を用いる．C_{b_i} は b_i が係りうる主辞の集合で，解析アルゴリズムと依存構造の制約によって決定される．文献 13) においては，単純な最大エントロピーモデルを用いて式 (2) を定式化しているが，本稿では複数仮説を用いることによる計算量の増加に対処するため，ゲート関数を用いた定式化を行った．定式化の違いに関わらず，最終的な係り受けの性能は同程度であった．

2.3 Sequential dependency analysis

Sequential Dependency Analysis (SDA) は文境界を表すメタシンボル ⟨b⟩ および将来的に観測される文節を表すメタシンボル ⟨c⟩ を導入することで文境界が未知である系列に対して依存構造解析を行う手法である[14]．SDA における依存構造解析は，非交差条件などの一般的な日本語係り受け解析の制約に従う．

本稿では，⟨b⟩ および ⟨c⟩ に加えて，主辞を持たない文節の係り先を擬似的に表現するメタシンボル ⟨ε⟩ を新たに導入する．⟨ε⟩ は主にフィラーなどの言い淀み (修飾辞と仮定) の主辞として使用されるため，言い淀み検出のために有効であると考えられる．

3. 文整形処理の定式化

3.1 定 式 化

本稿では，音声認識結果の整形を音声信号の系列 $\boldsymbol{o}$ が与えられた時に，対応する整形された書き起こし $\hat{\boldsymbol{w}}$ を発見する問題として定式化する．以下，文レベル情報を用いない場合と用いる場合の定式化について順に述べる．

3.1.1 文レベル情報を用いない場合

文レベル情報を使わない場合，$\hat{\boldsymbol{w}}$ は以下の式を用いて求められる[11]．⋆1

$$\hat{\boldsymbol{w}} = \operatorname*{argmax}_{\boldsymbol{w}} P(\boldsymbol{w})^{\alpha} \max_{\boldsymbol{s}} \prod_{i}^{N} \delta(w_i, s_i) P(s_i|\boldsymbol{o}), \tag{3}$$

ここで，$\boldsymbol{s}$ は音声信号系列 $\boldsymbol{o}$ に対応する発話そのものの書き起こし (音声認識結果) を表す．$\delta(w, s)$ は s が w に変換されうる場合には 1，そうでない場合には 0 となるような関数である．文献 15) と同様に，話し言葉とそれに対応する整形文のペアが利用できない場合を想定しているため，話し言葉の系列から整形文への変換モデルとして，パラレルコーパスから学習される統計モデルの代わりにルールベースのモデルを用いた．$P(s|\boldsymbol{o})$ は $\boldsymbol{o}$ に対する発話単語 s の事後確率であり，s は音響モデルと話し言葉の言語モデルを用いて作成されるコンフュージョンネットワークより得られる．$P(\boldsymbol{w})$ は $\boldsymbol{w}$ に対する書き言葉の言語モデルによる事前確率を表す．式 (3) の定式化では，コンフュージョンネットワークによって表現される仮説の上で max を求めることで，複数仮説を考慮した整形を行っている．この定式化は，コンフュージョンネットワークに仮説が含まれている限り，挿入，脱落，置換のあらゆる整形処理を扱うことが可能である．

3.1.2 文レベル情報を用いる場合

文整形処理のために文レベル情報を用いる目的は，文境界検出精度を高めるためであった

⋆1 本稿における定式化は文献 11) のものとは若干異なっている．本稿においては $\sum$ の代わりに max を使用し，β を使用していない．

が，文境界検出を行う過程で計算されるスコアも整形に有効であると考えられる．本稿で文境界検出に用いる improved-SDA が文境界を検出する過程においては，文節チャンキングスコア $P(\boldsymbol{b}|\boldsymbol{s})$ および依存構造解析スコア $P(\boldsymbol{d}|\boldsymbol{b})$ が計算される．これらのスコアを考慮することで，式 (3) は以下のようになる．

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, P(\boldsymbol{w})^{\alpha} \max_{\boldsymbol{s},\boldsymbol{b},\boldsymbol{d}} P(\boldsymbol{d}|\boldsymbol{b})^{\beta} P(\boldsymbol{b}|\boldsymbol{s})^{\gamma} \times \prod_{i}^{N} \delta(w_i, s_i|\boldsymbol{b}, \boldsymbol{d}) P(s_i|\boldsymbol{o}), \tag{4}$$

ここで，β および γ はそれぞれ $P(\boldsymbol{d}|\boldsymbol{b})$ および $P(\boldsymbol{b}|\boldsymbol{s})$ の重みである．Improved-SDA の結果 $\{\boldsymbol{b}, \boldsymbol{d}\}$ を考慮するために，話し言葉から整形文への変換ルールとして式 (3) における $\delta(w_i, s_i)$ の代わりに $\delta(w_i, s_i|\boldsymbol{b}, \boldsymbol{d})$ を用いている．

3.2 Improved-SDA を用いた変換ルール

文献 11) で使用した変換ルールを，improved-SDA の結果を考慮するために変更する．本稿で扱う変換ルールは，フィラーの削除および句点挿入であり，助詞の挿入は性能を低下させることがわかっているため考慮していない．

3.2.1 フィラー削除

フィラー削除のために以下のルールを使用する．

$$\delta(del, Filler\,; Filler \rightarrow \langle\varepsilon\rangle|\boldsymbol{b}, \boldsymbol{d}) = 1. \tag{5}$$

ここで *del* は単語の脱落を表す記号である．したがって，Improved-SDA によって，主辞が $\langle\varepsilon\rangle$ である，すなわち係り先が存在しなかったフィラーが削除される．

3.2.2 句点挿入

Improved-SDA による文境界検出結果を利用することで，より高精度な句点挿入を行うことができると考えられる．Improved-SDA による文境界検出結果を利用するために，以下の句点挿入ルールを使用する．

$$\delta(Period, \langle b\rangle|\boldsymbol{b}, \boldsymbol{d}) = 1, \tag{6}$$

ここで，$\langle b\rangle$ は improved-SDA によって検出された文境界を表すメタシンボルである．

4. Iterative decoding による複数仮説からの整形処理

複数仮説が，ラティスのようにある基準にしたがって束ねられており，N-best のように文として直接的に表現されていない場合，文全体に依存するような文レベル情報を適用することは困難である．したがって，そのような複数仮説を文レベル情報に基づいてリスコアリングすることは難しい．複数仮説が直接的に文として表現されている N-best を使用することで，そのような文レベル情報を使用することが可能となるが，N-best は数千の仮説を表現するために効率的な複数仮説の表現方法ではない．

この問題を解決するために，本稿では，コンフュージョンネットワークに含まれる仮説を文レベル情報を用いてリスコアリングするための手法である iterative decoding を利用することを提案する[12]．Iterative decoding を用いることで，コンフュージョンネットワークで表現される複数仮説に対して効率的に文レベル情報を適用できるようになる．本稿で提案する iterative decoding を用いた複数仮説から文レベル情報を用いて文整形を行うアルゴリズムを**図 1** に示す．

5. ベースライン

提案手法の有効性を示すために，提案法を以下に説明する 2 つの手法と比較する．

- フィラー除去
  フィラーは書き起こしの読みやすさに最も影響を与えると考えられる．そのため，書き起こしから品詞に基づいて自動的にフィラーを取り除いたものをベースラインとして使用した．本稿では，その品詞が “フィラー” または “感動詞” の場合に各単語をフィラーとして削除した．⋆1
- 1best 整形
  複数仮説の有効性を示すために，コンフュージョンネットワークから得られる認識結果の 1best 仮説のみを使用した整形を行った．1best 整形は，認識結果のコンフュージョンネットワークから 1best 仮説および脱落単語 *del* 以外の単語を全て取り除いたものに提案法を適用することで実現した．そのため，コンフュージョンネットワークから得られる単語の事後確率情報を使用している．
- 複数仮説を用いた整形
  文レベル情報の有効性を示すために，文レベル情報を用いずに複数仮説のみを用いた整形手法をベースラインとして使用する．文レベル情報を用いないため，句点は書き言葉コーパス (本実験では新聞) から学習される N-gram 統計量，すなわち，式 (3) における $P(\boldsymbol{w})$ に基づいて挿入される．

⋆1 形態素解析器には ChaSen + IPADic ver. 2.7.0 を使用した．

Input: an observation sequence $\boldsymbol{o}$
Output: a clean, readable transcript $\hat{\boldsymbol{w}}$
Variables:
$\boldsymbol{s}$, $\hat{\boldsymbol{s}}$, $\boldsymbol{s}'$: a spoken word sequence
$\boldsymbol{b}$, $\hat{\boldsymbol{b}}$, $\boldsymbol{b}'$: a *bunsetsu* sequence
$\boldsymbol{d}$, $\hat{\boldsymbol{d}}$, $\boldsymbol{d}'$: a dependency structure
$\boldsymbol{w}$, $\hat{\boldsymbol{w}}$, $\boldsymbol{w}'$: a clean, readable transcript
$CN = \{C_1, C_2, \ldots, C_{|CN|}\}$: a confusion network
$C_i = \{C_{i1}, C_{i2} \ldots, C_{i|C_i|}\}$: a set of words in the bin i
C_{ij} : the jth word in the bin i of a confusion network
x, $\hat{x}$: real value

```
1   begin
2     b̂ = φ, d̂ = φ, ŵ = φ
3     while the next utterance exists do
4       recognize the utterance and construct confusion \
          network and obtain confusion network CN
5       x̂ = 0.0
6       s' = ŝ, b' = b̂, d' = d̂, w' = ŵ
7       ŵ = φ
8       while ŵ ≠ ŵ̂ do
9         ŵ̂ = ŵ
10        for i = 1 to |CN|
11          for j = 1 to |C_i|
12            s = s' + GetHypothesis(CN, i, j)
13            b = b' + Chunking(s, b')
14            d = d' + Improved-SDA(b, d')
15            w = w' + Transform(s, b, d, w')
16            Compute score x by Eq. (4) using s, b, d, w
17            if x > x̂ then
18              ŝ = s, b̂ = b, d̂ = d, ŵ = w
19              x̂ = x
20              swap(C_i1, C_ij)
21            end if
22          end for
23        end for
24      end while
25    end while
26  end
```

図 1　Iterative decoding を用いた整形アルゴリズム．
Fig. 1　Transcription cleaning algorithm with iterative decoding.

6. 実　　験

6.1 実験条件

本実験で対象とするのは，大学院における講義を多数収録した CJLC コーパス[16] における講義 8 コマ分である．これらの講義の書き起こしに対して人手で整形文を作成し，正解整形文とした．認識結果と対応する整形結果の例を**図 2** に示す．また，テストデータの詳細を**表 1** に示す (Paraph. は人手による整形結果)．

表 1　テストセット諸元 (8 講義平均). *APP.* は adjusted perplexity，*OOV* は未知語率を示す．
Table 1　Statistics of the test sets (average of 8 lectures). *APP.* means adjusted perplexity. *OOV* indicates the ratio of OOV.

Duration (min.)	#Words		APP.		OOV [%]		#Filler [%]
	Manual	Paraph.	Manual	Paraph.	Manual	Paraph.	
67.6	11813	10192	182.6	159.7	3.5	3.9	7.2

音声認識のための音響モデルには，日本語話し言葉コーパス（CSJ）から学習した文脈依存のモデル（928 音節）を使用した．特徴量は MFCC，ΔMFCC，ΔΔMFCC，パワーの Δ および ΔΔ で計 38 次元である．各音節は left-to-right の 5 状態からなり，最終状態以外の 4 状態が出力分布を持つ．各分布は GMM で，混合数は 4，共分散行列はブロック型である．言語モデルには，話し言葉および書き言葉ともにトライグラムを使用し，話し言葉用の言語モデルは，CSJ の学会講演と模擬講演（2702 講演分），書き言葉のための言語モデルは，毎日新聞の記事のうち 91 年 1 月から 99 年 12 月までの 9 年分のデータから学習した（バックオフはウィッテンベル法を用いた）．語彙は，CSJ の学会講演と模擬講演に出現した上位 20000 万語に句読点を加えたものを使用した．実験の設定は基本的に文献 11) と同様である．認識結果および，作成したコンフュージョンネットワークの単語カバレッジと単語密度を**表 2** に示す．

Improved-SDA は，式 (2) に示す $P(b_i \rightarrow h_{b_i}|\boldsymbol{\Phi}(b_i, h_{b_i}, \boldsymbol{b}))$ の実現方法が異なることを除いて同様の構成で実装し，人手により節情報および依存構造が付与された CSJ のコア講演を用いて学習した．我々の実装は，$P(b_i \rightarrow h_{b_i}|\boldsymbol{\Phi}(b_i, h_{b_i}, \boldsymbol{b}))$ の実現方法の違いに関わらず，文境界検出と依存構造解析の精度において文献 13) とほぼ同等の性能を示した．

全てのパラメータは，CSJ に含まれる 4 つの講演からなる開発セットで決定した．式 (4) における α，β，γ はそれぞれ 0.15，0.0，5.0 となった．これは，文節チャンキングスコアは整形に有効であったが，依存構造解析のスコアは整形に有効でなかったことを意味する．これは，依存構造解析が音声認識誤りに対して脆弱であるためであると考えられる．

6.2 客観評価結果

提案法を評価するために，認識結果，ベースラインによる整形文，および提案法による整形文を人手による整形文とそれぞれ比較した．

実験結果を**表 3** に示す．表より，音声認識結果そのものを用いた場合が最も悪い性能を示していることがわかる (WER=64.2%)．音声認識結果からフィラーを取り除くことで性能の向上が見られた (64.2% → 58.4%†；ここで † はそれぞれ有意水準 1%で結果が有意であったことを意味する．)．1best 整形は，コンフュージョンネットワークから得られる認識結果

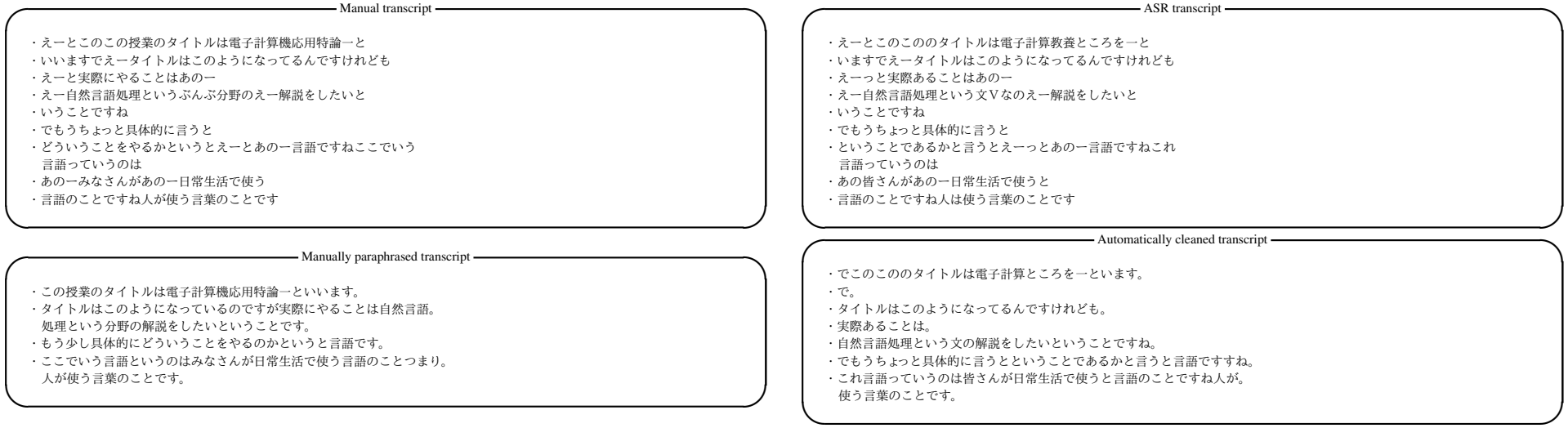

図 2　書き起こしの例.
Fig. 2　Examples of transcripts.

表 2　WER およびコンフュージョンネットワークの単語カバレッジと単語密度.
Table 2　WER, Word coverage and density of constructed confusion network.

正解 (ターゲット)	WER [%]	単語カバレッジ [%]	単語密度
人手書き起こし	43.8	80.9	3.83
人手整形文	64.2	75.9	

の事後確率と書き言葉の N-gram 情報を利用しながら複数仮説から整形文を選択することにより，フィラーのみならず言い淀みの削除が可能であるため，単純なフィラー除去に対して有意に改善した (58.4% → 55.2%†)．さらに複数仮説を利用することで，1best 仮説には出現しなかった単語を扱えるようになり，さらに改善した (55.2% → 54.2%†)．文レベル情報を用いることで，1best 整形および複数仮説はそれぞれ文レベル情報を用いない場合に比べて有意に改善した (55.2% → 53.8%†, 54.2% → 53.5%†)．文レベル情報を用いた 1best 整形に対し，さらに複数仮説を用いても有意に改善しなかったが (53.8% → 53.5%†)，複数仮説の利用は次に示す句点挿入精度の向上や，6.3 節に示すように理解し易さの向上に効果がある．

表 4 に句点挿入の評価結果を示す．句点挿入を評価するために，整形文と句点を含む人手

表 3　各書き起こしの評価結果 [%]
Table 3　Evaluation results of the transcripts [%]

Method	Del.	Ins.	Subs.	WER
認識結果	6.8	18.2	39.3	64.2
フィラー除去	8.7	11.0	38.7	58.4
1best 整形	16.1	7.3	31.8	55.2
+ 文レベル情報	18.7	5.0	30.1	53.8
複数仮説	15.4	7.0	31.7	54.2
+ 文レベル情報	16.7	5.7	31.1	53.5

による整形文のアライメントを取り，句点箇所の F 値を計算した．表中の発話単位は，音声認識に用いた処理単位の終端に句点を挿入する方法である．本稿において，音声認識に用いた処理単位は 200ms 以上のポーズに囲まれた音声区間である．表 3 の認識結果とフィラー除去による句点挿入は，発話単位に基づいて行なった．表より，発話単位による句点挿入が最も精度が悪いことがわかる (0.334)．1best 整形により，コンフュージョンネットワークからの事後確率と，N-gram の情報を用いた句点挿入を行うことで，発話単位による句点挿入

表 4　句点挿入結果の評価．
Table 4　Evaluation results of the insertion of periods.

Method	Recall	Precision	F
発話単位 (ポーズ)	0.562	0.252	0.334
1best 整形	0.475	0.463	0.467
+ 文レベル情報	0.457	0.543	0.489
複数仮説	0.487	0.421	0.450
+ 文レベル情報	0.497	0.517	0.501

表 5　被験者実験に使用したテストセット諸元．*APP.* は adjusted perplexity，*OOV* は未知語率を示す．
Table 5　Statistics of the test set for the subjective test. *APP.* means adjusted perplexity. *OOV* indicates the ratio of OOV.

Duration (min.)	#Words		APP.		OOV [%]		#Filler [%]
	Manual	Paraph.	Manual	Paraph.	Manual	Paraph.	
8.18	1654	1410	148.5	121.2	2.2	2.1	8.8

よりも大幅に精度が向上した (0.334 → 0.467)．1best 整形に対しさらに複数仮説を用いた場合，句点挿入精度が低下する結果となった (0.467 → 0.450)．これは，句点挿入を N-gram の情報に頼ることで，複数仮説を用いることによる湧き出し誤りの影響が大きくなったことが原因であると考えられる．文レベル情報を用いることで，1best 整形および複数仮説ともに句点挿入精度が改善した (0.467 → 0.489, 0.450 → 0.501)．特に，文レベル情報と複数仮説を同時に用いた場合に最高の性能となり (0.501)，文レベル情報と複数仮説を同時に用いる有効性を示している．

これら結果より，複数仮説とともに文レベル情報を用いることは文整形に有効であり，本稿で提案する iterative decoding を用いた手法はコンフュージョンネットワークで示される複数仮説を文レベル情報を用いて効果的にリスコアリング可能であるといえる．

6.3　主観評価結果

提案法による整形が音声認識結果の読み易さおよび理解し易さを改善したかどうかを評価するために，被験者 10 名による被験者実験を行った．本稿における読み易さおよび理解しやすさの定義は以下の通りである．

- 読み易さ：発話内容を知らない前提で，書き起こしが実際の意味を表しているかどうかに関わらず読み易かったかどうか．
- 理解し易さ：発話内容を知った前提で，実際の発話内容を理解し易かったかどうか．

被験者実験は，CJLC の 8 つの講義のうち 2 つの講義の 1 部分に対して行った．被験者実験に使用したデータの詳細を**表 5** に示す．また，認識結果および，コンフュージョンネットワークによる単語カバレッジと単語密度を**表 6** に示す．表 1 および表 2 と比較することで，被験者実験に使用したデータが実験データ全体の良いサンプリングであることがわかる．

被験者実験の手順は以下に示す通りである．

(1)　A と B の (整形された) 書き起こしを読む (それぞれ約 30 行，850 字，書き起こしは 4 つの小区間に分けられており (約 7 行) おおよそトピックに相当する).
(2)　それぞれの区間毎に，読み易さについて評価する (paired comparison).
(3)　正解書き起こしを読み，再び A と B を読む．
(4)　同様に，理解し易さについて評価する (paired comparison).

被験者は，両者の間に差を付けられない場合には “?” を使っても良いとした．句読点の存在によるバイアスを避けるため，句読点は全て取り除いて評価した．

被験者実験の結果を**表 7** に示す．表中の “count” は，各手法が選ばれた数を表しており，“†” および “††” はそれぞれ，有意水準 5%および 1%のもとで有意に上回っていたことを示す．

表より，フィラー除去は両基準において認識結果を有意に上回っていることがわかる．これより，フィラーは常に除去することが望ましいといえる．1best 整形 (文レベル情報無し) は，フィラー除去に対し，読み易さでは改善したが理解し易さは逆に低下する結果となった．これは，1best 整形が言い淀み箇所を削除することで読み易さを改善するが，言い淀み箇所の検出が十分に高精度でないために，内容の理解に不可欠な単語の脱落を生じ，理解し易さを低下させたものと考えられる．1best(文レベル情報無し) と 1best(文レベル情報有り) を比較すると，文レベル有りの場合が両基準において有意に改善しており，文レベル情報は整形に有効であるといえる．一方で，1best(文レベル情報無し) と複数仮説 (文レベル情報無し) を比較すると，複数仮説の利用は特に理解し易さに効果があったことがわかる．文レベル情報と複数仮説を同時に利用することで，複数仮説 (文レベル情報有り) は 1best(文レベル情報無し) を両方の基準で有意に改善した．一見，複数仮説を使用しない 1best(文レベル情報有り) で十分であるように見えるが，複数仮説 (文レベル無し) と 1best(文レベル情報無し) の比較結果より，複数仮説は 1best に存在しない仮説によって特に理解し易さを向上しており，両方の情報を同時に使用することでより良い結果を与えるといえる[⋆1]．

読み易さの観点から見ると，文レベル情報と複数仮説を同時に使用した複数仮説 (文レベル情報有り) は，単純なフィラー除去に対し，“フィラー除去 < 1best(文レベル情報無し) <

⋆1 表中で*付きの結果は被験者が異なるため，値の直接の比較には注意が必要である．

表 6 被験者実験に使用したテストセットに対する WER およびコンフュージョンネットワークの単語カバレッジと単語密度．
Table 6 WER, Word coverage and density of constructed confusion network of the test set for the subjective test.

Reference	WER [%]	単語カバレッジ [%]	単語密度
Manual	52.0	74.0	4.00
Paraphrased	69.0	70.1	

表 7 被験者実験結果．† および †† はそれぞれ有意水準 5%および 1%で統計的に有意に上回ることを示す．* は被験者が異なることを示す．
Table 7 Subjective test results. † and †† indicate statistical significance of the method under the significance level 0.05 and 0.01, respectively. * means that the subjects were different with other experiments.

評価基準	Method		Count		
	A	B	A	B	?
読み易さ	認識結果	フィラー除去 ††	8	70	2
	フィラー除去	1best 整形 (文レベル情報無し)††	26	50	4
	*1best 整形 (文レベル情報無し)	1best 整形 (文レベル情報有り)††	20	53	7
	*1best 整形 (文レベル情報無し)	複数仮説 (文レベル情報無し)	34	41	5
	1best 整形 (文レベル情報無し)	複数仮説 (文レベル情報有り)††	27	51	2
理解し易さ	認識結果	フィラー除去 ††	8	57	15
	フィラー除去 ††	1best 整形 (文レベル情報無し)	60	18	2
	*1best 整形 (文レベル情報無し)	1best 整形 (文レベル情報有り)††	4	65	11
	*1best 整形 (文レベル情報無し)	複数仮説 (文レベル情報無し)††	3	66	11
	1best 整形 (文レベル情報無し)	複数仮説 (文レベル情報有り)††	19	56	5
	フィラー除去	複数仮説 (文レベル情報無し)	38	28	14

複数仮説 (文レベル情報有り)” の関係から有意に改善しているといえる．しかし，理解し易さに関しては “フィラー除去 > 1best(文レベル無し)” であるから，提案法の優位性を示すことができない．しかしながら，フィラー除去と複数仮説 (文レベル情報無し) 間に統計的な有意差が無いことから，少なくともフィラー除去が複数仮説 (文レベル情報有り) を統計的に有意に上回ることはないといえる．提案法は，主として言い淀み箇所の脱落によって改善を得ているため，読み易さを有意に改善しながら，フィラー除去と同程度以上の理解し易さを保つことに成功しているといえる．

被験者実験による主観評価からも，文レベル情報と複数仮説を同時に用いることの有効性が示せたといえる．

7. おわりに

本稿では，音声認識結果の自動整形のために，文献 12) で提案される iterative decoding を利用して，コンフュージョンネットワークで表現される複数仮説と文レベル情報を同時に使用する手法を提案した．文レベル情報として improved-SDA から得られるチャンキングスコアおよび依存構造解析のスコアを利用することで文レベル情報を利用しない場合よりも高精度な整形を実現できた．また，improved-SDA による文境界検出結果を考慮することで，句点挿入精度を大幅に改善できた．被験者実験によっても，提案法の優位性が確認された．

リアルタイム性が要求され，コンフュージョンネットワークの構築が困難な場合には，複数仮説としてラティスを用いた方が良い．文献 17) において提案されるラティス上の仮説を文レベルの情報でリスコアリングするための手法を適用するなど，ラティスで表現される複数仮説と文レベル情報を同時に使用した整形手法の検討が今後の課題である．また，話し言葉の言い回しを書き言葉へ変換するルールの導入についても今後検討していきたい[9),10)]．

謝辞 本研究は文部科学省グローバル COE プログラム「インテリジェントセンシングのフロンティア」の支援を受けた．

参考文献

1) Fujii, Y., Yamamoto, K., Kitaoka, N. and Nakagawa, S.: Class Lecture Summarization Taking into Account Consectiveness of Important Sentences, Proc. Interspeech, pp.2438–2441 (2008).
2) Chelba, C. and Acero, A.: Indexing Uncertainty for Spoken Document Search, Proc. Interspeech, pp.61–64 (2005).
3) Togashi, S. and Nakagawa, S.: A browsing system for classroom lecture speech, Proc. Interspeech, pp.2803–2806 (2008).
4) Glass, J., T.J.Hazen, S.C., Malioutov, I., Huynh, D. and Barzilay, R.: Recent Progress in the MIT Spoken Lecture Processing Project, Proc. Interspeech, pp.2553–2356 (2007).
5) Kogure, S., Nishizaki, H., Tsuchiya, M., Yamamoto, K., Togashi, S. and Nakagawa, S.: Speech Recognition Performance of CJLC: Corpus of Japanese Lecture Contents, Proc. Interspeech, pp.1554–1557 (2008).
6) T.Kawahara, Y.Nemoto and Y.Akita: Automatic lecture transcription by exploiting presentation slide information for language model adaptation, Proc. IEEE-ICASSP, pp.4929–4932 (2008).
7) Shitaoka, K., Nanjo, H. and Kawahara, T.: Automatic transformation of lecture transcription

into document style using statistical framework, Proc. Interspeech, pp.2881–2884 (2004).
8) Hori, T., Willet, D. and Minami, Y.: Paraphrasing spontaneous speech using weighted finite-state transducers, Proc. SSPR, pp.210–222 (2003).
9) Neubig, G., Mori, S. and Kawahara, T.: A WFST-based Log-linear Framework for Speaking-style Transformation, Proc. Interspeech, pp.1495–1498 (2009).
10) Neubig, G., Akita, Y., Mori, S. and Kawahara, T.: Improved Statistical Models for SMT-Based Speaking Style Transformation, Proc. ICASSP, Dallas, Texas, USA (2010).
11) Fujii, Y., Yamamoto, K. and Nakagawa, S.: Improving the Readability of Class Lecture ASR Results using a Confusion Network, Proc. Interspeech, pp.3078 – 3081 (2010).
12) Deoras, A. and Jelinek, F.: Iterative Decoding: A Novel Re-scoring Framework for Confusion Network, Proc. ASRU, pp.282 – 286 (2009).
13) Oba, T., Hori, T. and Nakamura, A.: Improved Sequential Dependency Analysis Integrating Labeling-Based Sentence Boundary Detection, IEICE, Vol.E93-D, No.5, pp.1272–1281 (2010).
14) Oba, T., Hori, T. and Nakamura, A.: Sequential dependency analysis for online spontaneous speech processing, Speech Commn., Vol.50, pp.616–625 (2008).
15) Shungrina, M.: Formatting Time-Aligned ASR Transcripts for Readability, Proc. NAACL, pp.198 – 206 (2010).
16) 土屋，小暮，西崎，太田，山本，中川：日本語講義音声コンテンツコーパスの作成と分析，情報処理学会論文誌，Vol.50, No.2, pp.448–450 (2009).
17) Rastrow, A., Dreyer, M., Sethy, A., Ramabhadran, B. and Dredze, M.: HILL CLIMBING ON SPEECH LATTICES: A NEW RESCORING FRAMEWORK, Proc. ICASSP, pp.5032 – 5035 (2011).