整形された書き起こしからの整形・非整形部分の自動検出

太田健吾[†1]　土屋雅稔[†2]　中川聖一[†1]

話し言葉の正確な書き起こしを含む大規模コーパスは，様々な音声言語処理において必要不可欠である．しかし，話し言葉を正確に書き起こす作業は極めて高いコストを必要とするため，そのようなコーパスが実際に利用できるドメインは稀である．一方で，速記録や会議録は，幅広いドメインにおいて利用が可能である．ただし，速記録や会議録では，可読性を高めるために，フィラーや言い淀み，言い直しなどの話し言葉特有の現象は削除され，話し言葉特有の言い回しは適切な書き言葉に置き換えられていることが一般的である．このような背景を踏まえて，我々は，速記録や会議録(整形された書き起こし) を正確な書き起こしに半自動的に変換する枠組みの構築について検討してきた．この枠組みでは，まず整形された書き起こし中の整形箇所を自動検出し，検出された整形箇所のみを人手で書き起こすことにより，できるだけ小さいコストで，整形された書き起こしと正確な書き起こしのパラレルコーパスを作成する．本稿では，この枠組みのために，整形された書き起こしから整形箇所を自動検出する方法の改良法について述べる．最初に，整形された書き起こしと原音声のアラインメントを行い，次に，アラインメントによって得られた素性に基づくサポートベクターマシンを用いて整形箇所を検出する．後者の段階で，従来法とは異なり，音節を単位とする音響的素性を用いる．国会会議録を用いた評価実験の結果，実用的な検出精度を達成することができた．また，提案法によって抽出された非整形部分は，音響モデルの教師あり話者適応の発音ラベルとして有効であることを示した．

Automatic Detection of Edited Parts in Inexact Transcribed Corpora

Kengo Ohta,[†1] Masatoshi Tsuchiya[†2]
and Seiichi Nakagawa[†1]

Large-scale spontaneous speech corpora are crucial resource for various domains of spoken language processing. However, the available corpora are usually limited because their construction cost are quite expensive especially in transcribing speech precisely. On the other hand, inexactly transcribed corpora like shorthand notes, meeting records and closed captions are more widely available than precisely transcribed ones, because their inexactness reduces their construction cost. Because these corpora contain both precisely transcribed parts and *edited* parts, it is difficult to use them directly as speech corpora for learning acoustic models. Under this background, we have been considering to build an efficient semi-automatic framework for converting inexact transcripts to faithful ones or exact transcriptions. This paper describes a revised automatic detection method of edited parts in inexactly transcribed corpora for this framework. Our detection method consists of two steps: the first step is an automatic alignment between *edited* transcriptions and their utterances to discover the corresponding utterance for the certain edited transcription, and the second step is a support vector machine based detector of precise parts using several features obtained by the first step. Our experimental result shows that our proposed method achieves high accuracy of detecting precise parts, and shows that the partially-transcribed speech corpus generated by our proposed method is effective as training labels of lightly supervised speaker adaptation.

1. はじめに

各種の音声言語処理アプリケーションについて，大量の話し言葉の正確な書き起こしを含む大規模コーパスが利用できることは非常に重要である．たとえば，話し言葉を対象とする音声認識システムには，対象音声とドメインが一致し，かつ，話し言葉特有の言い回しに対応した言語モデルが不可欠である．そのような言語モデルを構築する最も単純な方法は，対象音声と同一ドメインの大規模な話し言葉コーパスから言語モデルを学習するという方法である．しかし，話し言葉を正確に書き起こす作業は極めて高いコストを必要とするため，あらゆるドメインに対して，そのようなコーパスが入手できると仮定することは現実的ではない．

それに対して，速記録や会議録は，正確な書き起こしより広く作成されており，比較的容易に入手が可能である．ただし，速記録や会議録では，可読性を高めるために，フィラーや言い淀み，言い直しなどの話し言葉特有の現象は削除され，話し言葉特有の言い回しは適切な書き言葉に置き換えられていることが一般的である．本稿では，このような書き起こしを，**整形された書き起こし**と呼ぶ．たとえば，国立国会図書館は，1947 年以降の全ての国

†1 豊橋技術科学大学 情報・知能工学系
Department of Computer Sciences and Engineering, Toyohashi University of Technology
†2 豊橋技術科学大学 情報メディア基盤センター
Information Media Center, Toyohashi University of Technology

ところ が ですね ， えー , この 資料 , 見 て み ます と 神奈川 県 の 場合 は ， け ，
結果 と し て 財政 的 に いー 豊か に なっ てる と .

(i) 正確な書き起こし

ところ が ， この 資料 を 見 て み ます と , 神奈川 県 の 場合 は ， 結果 と し て
財政 的 に 豊か に なっ ている .

(ii) 整形された書き起こし

図 1　正確な書き起こしと整形された書き起こしの例
Fig. 1　Example of precise and edited transcriptions

会の会議録を公開している[⋆1]．例を図 1 に示す．図 1 に見られるように，国会会議録では，可読性を重視するために，フィラー（例．”えー”, ”いー”）や言い直し（例．”け”），および冗長な言い回し（例．”ですね”, ”と”）はすべて削除されており，また，口語体の言い回し（例．”てる”）は文語体（例．”ている”）に置き換えられている．さらに，助詞の不足が補われている（例．”を”）ほか，一部の読点は速記者の判断によって追加あるいは削除されている．また，Williams らは，多数の非熟練作業者を利用して書き起こしを作成する方法を提案している[11]．この方法は，熟練作業者による書き起こしに比べて，作業コストの面では非常に安価ではあるものの，多数の書き起こしミスが含まれる欠点がある．このように，整形された書き起こしは，今後大量に利用できるようになる可能性が高い．

このような背景を踏まえて，我々は，整形された書き起こしを正確な書き起こしに半自動的に変換する枠組みの構築について検討してきた．この枠組みでは，まず整形された書き起こし中の整形箇所を自動検出し，検出された整形箇所のみを人手で書き起こすことにより，できるだけ小さいコストで，整形された書き起こしと正確な書き起こしのパラレルコーパスを作成する．先に述べた通り，整形された書き起こしは，話し言葉的現象を削除 · 整形して書き言葉に近づけたテキストであるから，このパラレルコーパスは，話し言葉特有の言い回しから書き言葉特有の言い回しへの変換パターンの構築[6] に利用可能である．このようにして構築された言い回し変換パターンと，整形された書き起こしを対象としてフィラーおよびポーズを復元する手法[14),16)] を組み合わせることにより，任意の書き言葉コーパスから，そのコーパスのドメインについての話し言葉に対応した言語モデルを自動構築することが，本研究の目標である．また，本手法により，整形されていない部分，すなわち正確に書き起こされた部分も同時に検出されることから，これらを音響モデルの教師あり学習[4] に用いることも可能である．

本稿では，この枠組みのために，整形された書き起こしから整形箇所を自動検出する方法の改良法を提案する．本手法は，2 つのステップからなる．第 1 に，整形された書き起こしとその原音声とでアラインメントを行い，第 2 に，アラインメントによって得られた素性に基づく Support Vector Machine (SVM) を用いて，整形箇所を検出する．第 2 段階において，従来法[7),15)] とは異なり，音節を単位とする音響的素性を用いる．

テキストと音声とのアラインメントの応用例として，たとえば Lamel ら[4] は，音響モデルの教師あり学習において，信頼できない学習データを除去するためにアラインメントを利用している．また，Roy ら[9] は，アラインメントから得られた音響スコアに基づいて，対象音声の書き起こし作業の難しさを推定している．さらに，丸山ら[13] は，ドキュメンタリー番組を対象として，字幕の送出タイミングを検出するためにアラインメントを利用している．これに対し，我々の提案法では，整形された書き起こしと原音声とでアラインメントを行うため，両者の不一致部分（＝整形箇所）では，アラインメントにおける音響スコアが低下し，また，各音節の対応区間にばらつきが生じる．この音響スコアの低下や音節区間のばらつきを手がかりとして，整形箇所を自動検出することが提案法のねらいである．

本稿の構成は以下の通りである．まず，2 節では，整形 · 非整形部分の検出方法について述べる．3 節では，国会会議録を用いた実験によって検出法を評価する．最後に，4 節で本研究のまとめと今後の課題について述べる．

2. 整形 · 非整形部分の検出

本稿では，整形 · 非整形部分の検出を 2 通りの方法で行う．第 1 の方法は，我々が従来から提案している方法[7),15)] で，強制アラインメントと Support Vector Machine (SVM) を用いて整形箇所を検出する．第 2 の方法は，Paulik らによって提案されている方法[8)] で，大語彙連続音声認識 (LVCSR) の認識結果を用いる方法である．以下では，それぞれの方法について詳細を述べる．

2.1 強制アラインメントと SVM による整形 · 非整形部分の検出

我々が従来から提案している検出法[7),15)] は，2 段階からなる．第 1 に，整形された書き

⋆1 http://kokkai.ndl.go.jp/

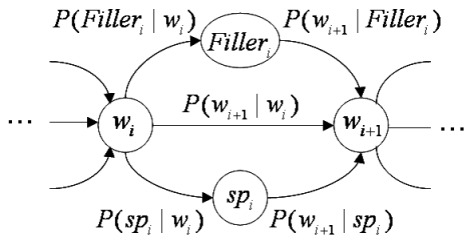

図 2　2-gram 言語モデルに基づく制約
Fig. 2　2-gram constraint for automatic alignment

起こしと原音声を強制アラインメントする．第 2 に，強制アラインメントによって得られた素性に基づいて動作する SVM を用いて整形箇所を検出する．以下，それぞれの段階について詳細を述べる．

2.1.1　整形された書き起こしと原音声の強制アラインメント

提案法の第 1 段階として，整形された書き起こしとその原音声との強制アラインメントを行う．最初に，整形された書き起こしの単語列 $w_1 w_2 \cdots w_n$ から**図 2** のような 2-gram 言語モデルを作成する．図 2 において，w_i は整形された書き起こしに含まれる i 番目の単語であり，$Filler_i$ および sp_i は w_i の直後に出現するフィラーとショートポーズである．明らかに，図 2 の言語モデルは，整形された書き起こし，または単語間にフィラー · ポーズが挿入された単語列のみを受理する．よって，図 2 の言語モデルを用いて原音声の連続音声認識を行うと，単語間にフィラーとショートポーズの挿入を許すという制約のもとで，整形された書き起こしに含まれる各単語を原音声にアラインメントできる[⋆1]．

図 3 に，正確な書き起こしと原音声のアラインメント例，および整形された書き起こしと原音声のアラインメント例を示す．図 3 に見られるように，整形された書き起こしと原音声とのアラインメントを行うと，両者の不一致部分（すなわち，整形箇所）において，実際の発声とは異なるモデルが強制的に対応付けられることになる．この結果，周囲の音節の対応区間が歪められ，また，モデルの不一致によって音響スコアが低下する．図 3 の例では，音節/*si*/の区間が不適切に延びており，また，モデル/*te*/が音節/*ne*/に，ショートポーズモデル/*sp*/が音節/*te de su*/にそれぞれ強制的に対応付けられている．よって，音節の区間が極端に長い/短い部分や，音響スコアが通常よりも低い部分は，整形箇所である可能性が高いといえる．

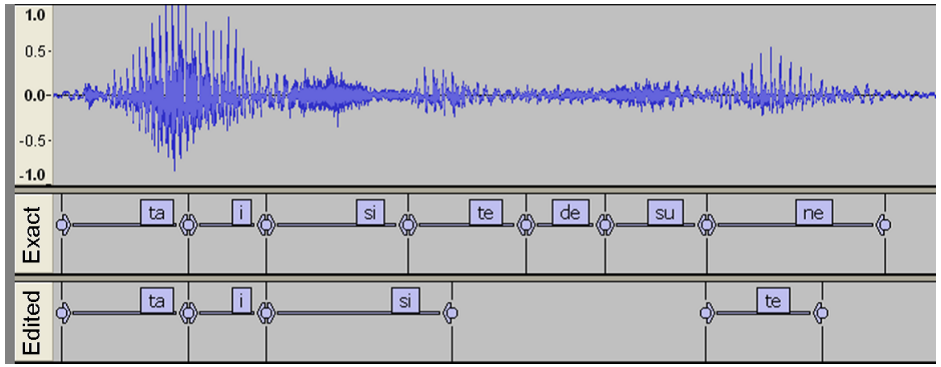

図 3　正確な書き起こし/整形された書き起こしと原音声とのアラインメント例
Fig. 3　Example of alignment between edited transcription and its corresponding utterance

なお，多くの整形された書き起こしにおいて，書き起こし中の各文と原音声中の各発話の対応に関する情報は付与されていないことが多い．これを考慮し，我々は，連続音節認識結果の音節数に基づいて，各発話の単語数を推定することにより，両者の対応付けを自動で行った．この対応付けに基づいて，書き起こし区間に対し，原音声の発話区間を切り出した上で，書き起こし中の各文とアラインメントを行う．予備実験により，このように各発話の単語数を自動で推定した場合でも，各発話の単語数が正確に得られた場合と同等の精度でアラインメントが行えることを確認している．

2.1.2　SVM による整形 · 非整形部分の検出

提案法の第 2 段階として，2.1.1 節の強制アラインメントによって得られた素性に基づく SVM を用いて，整形された書き起こしから整形箇所を検出する．本研究では，整形箇所の検出を，整形された書き起こしに含まれる各単語に対する二値分類問題として定式化する．つまり，整形された書き起こしに含まれる各単語を，整形された単語か，整形されていない単語かの 2 種類に分類する． SVM のツールキットとしては，TinySVM（ver 0.09）[3] を用いる．カーネル関数には多項式カーネルを採用する．

本稿では，各々の単語について，以下の特徴量を用意する．

- 音響的素性
  - 単語単位の音響尤度
  - 音節単位の音響的素性

⋆1 従来法[15] では，ショートポーズの挿入のみを考慮し，フィラーの挿入を考慮していなかった．

- 言語的素性
  - 単語
  - 品詞
  - 単語に含まれる音節数

判定対象となる単語，その直前 2 単語，その直後 2 単語に関する上記の特徴量を組み合わせて，判定用の素性として用いる．素性の選択にあたっては，Huang らの文献[2] を参考にした．以下では，音響的素性について詳細を述べる．

2.1.2.1 単語単位の音響尤度

強制アラインメントによって得られた音響尤度を素性として利用する．ただし，この音響尤度（対数尤度）は単語の時間長によって正規化し，さらに，連続音節認識によって得られた音響スコア（対数尤度）との差分を取る．これは，対数事後確率を素性として用いることに相当する．

2.1.2.2 音節単位の音響的素性

従来法[7),15)] では，整形 · 非整形を判定しようとしている単語を単位として，その単語に含まれる全ての音節の音節長の平均や分散を素性として用いていた．しかし，単語に含まれる全ての音節に基づく素性は，音節単位での異常さを検出するには不十分な場合がある．例として，国会会議録では「ところが」と書き起こされている部分について，原音声を正確に書き起こすと「ところがですね」となっていて，以下のようにアラインメントされた場合を考える．

国会会議録 (整形された書き起こし)	to	ko	ro	ga
正確な書き起こし	to	ko	ro	ga-de-su-ne

この場合，最初の 3 音節/*to*/ /*ko*/ /*ro*/の音響尤度や音節長には一切の異常はなく，最後の音節/*ga*/の音響尤度と音節長のみに異常が現れる．そのため，単語に含まれる全ての音節の音節長の平均や分散という素性では，異常さが 1/4 に薄められてしまい，整形部分を取り出す素性として使いにくくなることが予想される．この問題を回避するには，単語に含まれる全ての音節に基づく素性の代わりに，単語に含まれる音節から最も異常と推測される音節を 1 つ代表として取り出し，その代表音節に基づく素性を用いる方法が有効と考えられる．本稿では，代表音節を取り出す尺度として，**正規化音節長**と**正規化音響尤度**という 2 つの尺度を用いる．

正規化音節長の定義は，以下の通りである．最初に，国会会議録の原音声を正確に書き起こした音節列 s_1^N と原音声をアラインメントし，各音節の音節長 $d(s_i)$ を求める．次に，音節の種類 x 毎に，音節長の平均 $E_d(x)$ と分散 $V_d(x)$ を求める．

$$E_d(x) = \frac{\sum_{i=1}^{N} \delta(s_i = x) d(s_i)}{\sum_{i=1}^{N} \delta(s_i = x)} \tag{1}$$

$$V_d(x) = \frac{\sum_{i=1}^{N} \delta(s_i = x) \left(d(s_i) - E_d(x)\right)^2}{\sum_{i=1}^{N} \delta(s_i = x)} \tag{2}$$

実際に出現した音節 s_j の音節長について，平均 $E_d(s_j)$ と分散 $V_d(s_j)$ を用いて，次式のように平均 0，分散 1 に正規化する．

$$\tilde{d}(s_j) = \frac{d(s_j) - E_d(s_j)}{\sqrt{V_d(s_j)}} \tag{3}$$

このように正規化した音節長 $\tilde{d}(s_j)$ を，**正規化音節長**と呼ぶ．正規化音節長は，音節の種類毎に自然な音節長があり，かつ，音節長は正規分布に従うという仮定の下で，実際に出現した音節が，どの程度に異常な音節長となっているかの尺度として使うことができる．

正規化音響尤度の定義は，以下の通りである．最初に，国会会議録の原音声を正確に書き起こした音節列 s_1^N と原音声をアラインメントし，各音節の音響尤度 $L(s_i)$ を求める．次に，音節の種類 x 毎に，音響尤度の平均 $E_L(x)$ と分散 $V_L(x)$ を求める．

$$E_L(x) = \frac{\sum_{i=1}^{N} \delta(s_i = x) L(s_i)}{\sum_{i=1}^{N} \delta(s_i = x)} \tag{4}$$

$$V_L(x) = \frac{\sum_{i=1}^{N} \delta(s_i = x) \left(L(s_i) - E_L(x)\right)^2}{\sum_{i=1}^{N} \delta(s_i = x)} \tag{5}$$

次に，実際に出現した音節 s_j の音響尤度について，平均 $E_L(s_j)$ と分散 $V_L(s_j)$ を用いて，次式のように平均 0，分散 1 に正規化する．

$$\tilde{L}(s_j) = \frac{L(s_j) - E_L(s_j)}{\sqrt{V_L(s_j)}} \tag{6}$$

このように正規化した音響尤度 $\tilde{L}(s_j)$ を，**正規化音響尤度**と呼ぶ．正規化音響尤度は，正規化音節長と同様に，音節の種類毎に自然な音響尤度があり，かつ，音響尤度は正規分布に従うという仮定の下で，実際に出現した音節が，どの程度に異常な音響尤度となっているかの尺度として使うことができる．

国会会議録においては，言い直し · 言い淀みの削除が最も多い整形処理であることを考慮すると，正規化音節長が最大であるような音節，または，正規化音響尤度が最小であるような音節が，その単語内において最も異常な音節である可能性が高いと考えられる．そのため，以下の素性を用いる．

（1） 当該単語中の正規化音節長の最大値
（2） 当該単語中の正規化音響尤度の最小値
（3） 当該単語中で正規化音節長が最大な音節の正規化音響尤度
（4） 当該単語中で正規化音響尤度が最小な音節の正規化音節長

2.2 大語彙連続音声認識結果による整形 · 非整形部分の検出

Paulik らは，整形された書き起こしと LVCSR の認識結果を，音声認識精度を求める場合と同じ方法を用いて単語単位でアラインメントを行い，人手で与えた閾値以上に連続して一致した部分を発音ラベルとして用いて話者適応する方法を提案している[8]．本稿では，Paulik らの方法を利用して，以下の基準により整形 · 非整形部分の検出を行う．

（a） 閾値以上に連続して一致した部分を，非整形部分とする．
（b） 非整形部分以外の全てを，整形部分とする．

2.3 整形 · 非整形部分の検出性能の目標値

本稿では，整形箇所の検出性能として，精度 33%，再現率 50%を目標値とする．ここで例として，100 単語からなる会議録があり，この内の 10 単語が整形箇所である場合を考える．この場合，目標値通りの性能が達成できたとすると，15 単語が整形箇所として検出され，この内の 5 単語が真の整形箇所である．従って，書き起こし作業者は，計 30 単語のみを確認することで，10 単語の整形箇所を見つけることができる．これは，100 単語すべてを確認するのと比べ，約 3 倍以上の作業効率である．

また，本手法を逆に適用して，非整形箇所，すなわち正確な書き起こし部分を検出することも可能である．上述の場合，100 単語の会議録の内，90%が非整形箇所である．これに対し，本手法を適用すると，85 単語が非整形箇所として検出され，この内の 80 単語，すなわち 94%が真の非整形箇所である．従って，より正確な書き起こしを抽出できたことになる．

3. 評 価 実 験

3.1 実 験 条 件

実験には，国会会議録の一部を用いた．実験データの諸元を表 1 に示す．なお，テストコーパスの話者 11 名の内 1 名は女性のため，実験結果に悪影響を与えている可能性がある．

表 1 実験データ諸元
Table 1 Data statistics

	学習	テスト
時間長 (min)	42	60
話者数	7	11
単語数	7.2k	10.8k
整形単語数	604	426
整形単語率 (%)	8.4	3.9

表 2 音響分析条件
Table 2 Conditions of acoustic analysis for input speech

サンプリング周波数	16kHz
プリエンファシス	0.98
分析窓	Hamming 窓
分析窓長	25ms
フレームシフト	10ms
特徴パラメータ	MFCC + △MFCC + △△MFCC + △Pow + △△Pow (38 dimensions)

提案法 (2.1 節) の強制アラインメントおよび連続音節認識を行うデコーダには，SPOJUS++[1] を用いた．音響モデルは，CSJ[5] から学習した音節モデル（left-to-right 型 HMM, 5 状態 4 出力分布，単混合全共分散行列，男性話者モデル）を用いた．モデル数は 116 である．音響分析条件は**表 2** の通りである．予備実験より，図 2 の各状態の遷移確率として，$P(Filler_i|w_i) = 0.05$, $P(w_{i+1}|w_i) = 0.475$, $P(sp_i|w_i) = 0.475$ という値を用いた．

LVCSR の認識結果を用いる方法 (2.2 節) のデコーダとしても，SPOJUS++を用いた．音響モデルは，CSJ から学習した左コンテキスト依存音節モデル（left-to-right 型 HMM, 5 状態 4 出力分布，64 混合対角分散行列，男性話者モデル）を用いた．この左コンテキスト依存音節モデルは，116 種類の音節に対して 8 種類の先行音素 (/a/, /i/, /u/, /e/, /o/, /N/, /文頭/, /促音/) をコンテキストとして考慮しているため，モデル数は 928 である．音声分析条件は表 2 の通りである．言語モデルは，国会会議録に収録された 38,668K 語の整形された書き起こし (会議数としては 1,083 件に相当) から構築した形態素 3-gram モデルを用いた．国会会議録はフィラーおよびポーズが削除されているため，CSJ から学習したフィラー予測モデルおよびポーズ予測モデルを用いて，フィラーおよびポーズを考慮した言語モデルを作成した[14),16)]．スムージングには，一般的な Witten-Bell バックオフを用いた．

3.2 整形部分の自動検出

図 4 に，提案法による整形部分の自動検出結果を示す．図 4 の実線は，音響的素性と言

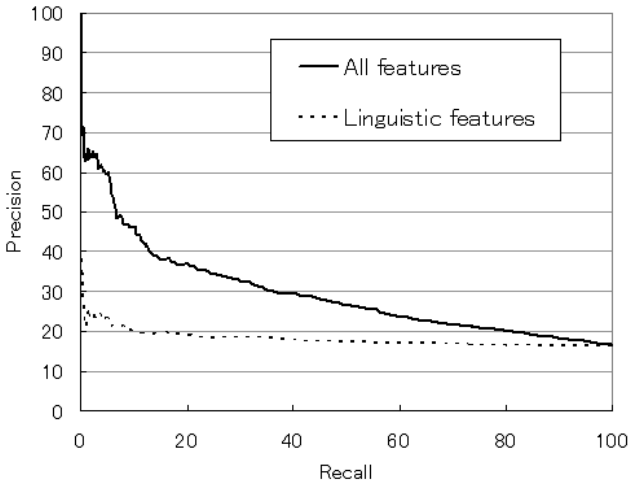

図 4　提案法による整形部分検出
Fig. 4　Detection of edited parts using the proposed method

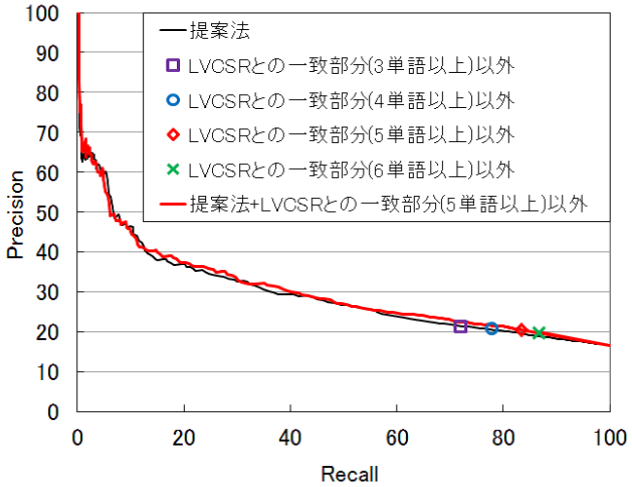

図 5　大語彙認識結果と組み合わせた整形部分検出
Fig. 5　Detection of edited parts using LVCSR combination

語的素性の両方を使って整形部分を自動検出した場合の結果，図 4 の破線は，言語的素性のみを使って整形部分を自動検出した場合の結果である．この 2 つの結果の違いから，音響的素性は，整形部分の自動検出に対して効果的であることが分かる．また，図 4 より，正確に整形部分を検出することは困難だが，2.3 節で述べた性能の目標値は達成できているから，パラレルコーパス作成の作業効率は 2.3 節の想定通りに改善できると考えられる．

図 5 に，提案法，LVCSR の認識結果を用いる方法，および両者を併用した結果を示す．両者の併用とは，以下のような方法である．

(a)　整形された書き起こしと LVCSR の認識結果が，閾値以上に連続して一致した部分を，非整形部分とする．

(b)　それ以外の部分を対象として，提案法による判定結果を採用する．

国会会議録の場合は，LVCSR の認識精度が比較的低いため[⋆1]，LVCSR の認識結果を用いる方法による性能は，提案法とほとんど変わらないことが分かる．そのため，両者を併用する方法についても，ほとんど変わらない性能となっている．

⋆1 **表 3** の「話者適応なし」行を参照．

3.3　非整形部分の自動検出

整形された書き起こしから原音声と一致する部分 (すなわち，非整形部分) を取り出す実験を行った結果を**図 6** に示す．図 6 より，整形された書き起こしに含まれている非整形部分は，音節単位で 80.1%である．よって，提案法を用いず，整形された書き起こし全体を非整形部分と見なすと，非整形部分の自動検出精度は 80.1%になる．提案法により，整形された書き起こしの 60%を選択すると，非整形部分の自動検出精度は 86.5%まで改善された．なお，この結果は，単語単位では 83.7%から 88.9%への改善に相当する．また，図 6 より，LVCSR の認識結果を用いる方法は，特に一致部分の長さの閾値が小さい場合には，提案法とほとんど性能が変わらないことが分かる．そのため，提案法と LVCSR の認識結果を用いる方法を併用した場合については，再現率が 40% ～ 60% の区間において，やや改善効果が見られる．

3.4　自動検出された発音ラベルを用いた話者適応

この節では，自動検出された非整形部分を発音ラベルとして用いて話者適応を行った実験結果について述べる．

HTK HMM toolkit ver 3.4.1[12] を利用し，MAP[10] による話者適応を行う．音声認識用

表 3　話者適応実験
Table 3　Results of Speaker Adaptation

話者適応用発音ラベルの作成方法	話者適応用発音ラベル諸元			音声認識性能	
	ラベル数	精度 (%)	カバー率 (%)	Cor. (%)	Acc. (%)
話者適応なし	0	—	0	71.5	67.2
国会会議録全体を発音ラベルとして使用	5,155 (91.4%)	80.1	84.3	74.3	70.6
国会会議録と LVCSR の一致部分 (5 単語以上) を使用	1,728 (30.6%)	90.0	69.3	73.9	70.1
提案法によって取り出した発音ラベル ($recall = 90\%$) を使用	4,056 (71.9%)	83.0	83.0	75.1	70.7
提案法によって取り出した発音ラベル ($recall = 90\%$) と，国会会議録と LVCSR の一致部分 (5 単語以上) の併用	4,212 (74.7%)	83.5	83.2	75.0	71.0
正確な書き起こしの音節ラベルを使用	5,642 (100.0%)	100.0	92.6	76.2	71.6

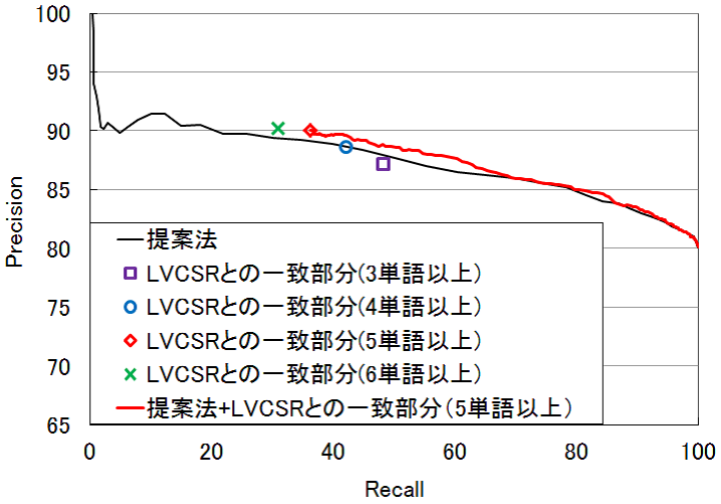

図 6　非整形部分の検出 (音節単位)
Fig. 6　Detection of precise parts (syllable units)

デコーダ，音響モデルおよび言語モデルは，LVCSR の認識結果に基づいて整形 · 非整形部分を検出する時に用いたもの (詳細は，3.1 節を参照) と同じである．

表 3 に，話者適応の実験結果を示す．なお，音響モデルの話者適応用学習コーパスは，男性話者 3 名の 5,642 音節からなり，テストコーパスは同じ話者 3 名の 4,411 音節からなる．表 3 の「精度」列は，学習用データに含まれる正しい発音ラベルの比率を，「カバー率」列は，テスト用データに含まれる音節が学習用データに含まれる比率 (token 単位) を示す．「国会会議録全体を発音ラベルとして使用」行は，提案法を用いずに，国会会議録 (整形された書き起こし) 全体を発音ラベルとして用いて話者適応を行った結果である．表 3 の「提案法によって取り出された発音ラベルを使用」行は，提案法を用いて $recall = 90\%$ の条件で取り出した非整形部分を発音ラベルとして用いて話者適応を行った結果である．「正確な書き起こしの音節ラベルを使用」行は，原音声を正確に書き起こした発音ラベルを用いて話者適応を行った結果であり，この学習コーパスに基づいて話者適応を行った場合の性能の上限を示す．表 3 より，提案法は，国会会議録全体を利用する方法よりも話者適応に用いる発音ラベル数が少ないにも関わらず，ベースライン手法 (国会会議録全体を利用する方法) よりも話者適応後の音声認識性能が良いことが分かる．これは，提案法によって取り出された発音ラベルが，話者適応に対して有効であることを示している．国会会議録と LVCSR による認識結果の一致部分 (5 単語以上) の発音ラベルを使用する方法は，話者適応用発音ラベルの精度は高いものの，カバー率が低いため，話者適応の結果が悪くなっている．また，図 6 より，提案法と，提案法と国会会議録と LVCSR による認識結果の一致部分 (5 単語以上) を併用する方法の 2 つは，$recall = 90\%$ の条件では，非整形部分検出の性能がほとんど同じである．そのため，話者適応の結果も若干の改善にとどまった．

4. おわりに

本稿では，整形された書き起こしから整形箇所を自動検出する手法を提案した．提案法は 2 段階からなる．第 1 に，整形された書き起こしと原音声のアラインメントを行い，第 2 に，アラインメントによって得られた素性に基づく SVM を用いて，整形箇所を検出する．国会会議録を用いた評価実験により，提案法は，整形された書き起こしと正確な書き起こし

のパラレルコーパスをできるだけ少ないコストで用意するための目標性能を達成できることを示した．また，提案法によって取り出された非整形部分は，話者適応の発音ラベルとして有効であることを示した．

提案法 (または，提案法と LVCSR の認識結果を用いる方法の併用) を用いて非整形部分を取り出す場合，非整形部分の精度を高くすると，再現率が低くなる (図 6)．そのような非整形部分を発音ラベルとして用いて話者適応する場合，発音ラベルの質は良いが，カバー率が低いため，話者適応の効果は小さくなる．より大規模なコーパスを用意することによって，カバー率の低下を補えるか検討することは，今後の課題である．また，正確な書き起こしの構築作業をどの程度効率化できるかを，被験者実験によって評価する予定である．

参 考 文 献

1) Fujii, Y., Yamamoto, K. and Nakagawa, S.: Large Vocabulary Speech Recognition System: SPOJUS++, *Proceeding of 11th WSEAS International Conference MUSP–11* (2011).
2) Huang, X., Acero, A., Acero, A. and Hon, H.: *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR (2001).
3) Kudoh, T.: TinySVM. `http://chasen.org/~taku/software/TinySVM/`.
4) Lamel, L., Gauvain, J.L. and Adda, G.: Investigating lightly supervised acoustic model training, *Processing of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp.477–480 (2001).
5) Maekawa, K.: Corpus of Spontaneous Japanese: Its Design and Evaluation, *Proceeding of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, pp.7–12 (2003).
6) Neubig, G., Akita, Y., Shinsuke, M., and Tatsuya, K.: Improved Statistical Models for SMT-based Speaking Style Transformation, *Proceeding of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp.5206–5209 (2010).
7) Ohta, K., Tsuchiya, M. and Nakagawa, S.: Detection of Precisely Transcribed Parts from Inexact Transcribed Corpus, *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, pp.541–546 (2011).
8) Paulik, M. and Panchapagesan, P.: Leveraging Large Amounts of Loosely Transcribed Corporate Videos for Acoustic Model Training, *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, pp.95–100 (2011).
9) Roy, B.C., Vosoughi, S. and Roy, D.: Automatic Estimation of Transcription Accuracy and Difficulty, *Proceeding of Interspeech*, pp.1902–1905 (2010).
10) Tsurumi, Y. and Nakagawa, S.: An Unsupervised Speaker Adaptation Method For Cotinuous Parameter HMM By Maximum A Posteriori Probability Estimation, *Proceedings of ICSLP'94*, pp.431–434 (1994).
11) Williams, J.D., Melamed, I.D., Alonso, T., Hollister, B. and Wilpon, J.: Crowd-Sourcing for Difficult Transcription of Speech, *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, pp.535–540 (2011).
12) Young, S.J., Evermann, G., Gales, M. J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P.C.: *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK (2006).
13) 丸山一郎，阿部芳春，江原暉将，白井克彦：ドキュメンタリー番組における字幕送出タイミング検出の一検討，日本音響学会秋季研究発表会講演論文集，3–Q–30, pp.177–178 (1999).
14) 太田健吾，土屋雅稔，中川聖一：フィラー予測モデルに基づく話し言葉言語モデルの構築，情報処理学会論文誌， Vol.50, No.2, pp.477–487 (2009).
15) 太田健吾，土屋雅稔，中川聖一：整形された会議録とその原音声のアラインメントに基づく整形箇所の自動検出，第 5 回音声ドキュメント処理ワークショップ講演論文集 (2011).
16) 太田健吾，土屋雅稔，中川聖一：ポーズを考慮した話し言葉言語モデルの構築，情報処理学会論文誌， Vol.53, No.2 (2012). (accepted).