非語彙的な表現を利用した音声対話の節境界同定の検討

牧本慎平† 柏岡秀紀†‡§ ニック キャンベル†‡§
† 奈良先端科学技術大学院大学 情報科学研究科
‡ 情報通信研究機構 知識創成コミュニケーション研究センター
§ 国際電気通信基礎技術研究所 音声言語コミュニケーション研究所
{shimpei-m, kashioka, nick}@is.naist.jp

1　はじめに

音声言語を処理する上で，発話をどのような単位で区切るかは重要な問題である．発話を意味的に適切な単位で区切ることによって，書き起こしテキストの可読性が向上するだけでなく，音声自動翻訳などのアプリケーションにおいても効率的に精度の高い結果を得ることが期待できる．独話における節境界・文境界同定の研究は日本語においてもこれまで成されてきた [8, 11]．また，英語の独話・対話における音響的特徴等を用いた統計的機械学習ベースの節同定の研究も報告されている [3, 4]．しかしながら，日本語音声対話における処理単位は発話ターンや発話内のポーズなどをもとにしたものが中心であった．これらの処理単位は，比較的フォーマルな状態での音声対話を対象としたものであり，雑談などのラフな対話においては境界を定めるのが難しい．そこで本稿では，教師ありの統計的機械学習をベースとした日本語音声対話における節境界同定の手法を提案する．

本稿で，我々が注目したのはフィラーや笑い，言い淀みなどの非語彙的な表現である．音声発話，特に対話などのインタラクションがある環境ではこれらの表現は頻繁に登場する．これらは従来，音声言語を処理する上での性能を低下させるノイズとして扱われてきた．しかしながら，一般的に人間の音声対話ではこれらの表現は円滑なコミュニケーションをする上での重要な要因となっている．特に，節境界同定に関して考えるなら，これらの表現の一部が，談話の構造を形作る discourse marker [6] として機能していることがあげられる [10]．そのため，これら非語彙的表現の情報は節境界同定のタスクにおいて，強く貢献できるものであると考えられる．

我々はこれまでに，相手話者の発話状態などの情報を用いた統計的機械学習の手法によって，対話内の非語彙的な表現を抽出する手法について提案した [12]．本稿では，これら非語彙的な表現を用いることによって，音声対話内の節境界の同定の性能向上にどれだけ貢献できるかを検証する．

本稿では，2 節において，我々が用いる自由対話コーパス ESP_C とそれに施したアノテーションについて言及し，3 節で本稿にて提案する非語彙的表現を用いた機械学習による節境界同定の手法について説明し，続く 4 節でその評価実験を行なった．最後に 5 節でまとめと今後の課題について述べる．

2　音声対話コーパスとアノテーション

本節では，我々が対象とした対話コーパス ESP_C とそれに施した非語彙的な表現と節境界情報のアノテーションについて述べる．

2.1　ESP_C コーパス

ESP_C コーパスは JST/ATR Expressive Speech Corpora [2] のサブセットであり，電話による 2 話者の対話が収録されている．コーパスの書き起こしに含まれている情報は，発話者，発話開始時間，発話時間，発話内容である．図 1 は ESP_C コーパス書き起こしテキストの一部である．

```
JFA-JFB-322.625-1.754  うーんだからすごく若い
JFB-JFA-324.679-1.104 あそうハハハ
JFA-JFB-324.735-3.498 エハハハハハ
JFB-JFA-326.240-2.547 じゃ話の内容も若返るんじゃ
ないですか
JFA-JFB-328.298-0.548 ヒーハ
JFA-JFB-328.861-1.337 ハーちょっとエネルギーを
JFA-JFB-330.210-0.478 アハア
JFB-JFA-330.423-0.933 ハハハ
JFA-JFB-330.708-0.206 ハッ
JFA-JFB-330.925-1.074 いただこうかと
```

図 1　ESP_C 書き起こしの一部

ESP_C コーパスは日本語による内容自由の対話であり，1 回のセッションは約 30 分で構成されている．参加者は全 10 名であり，日本語を母語としない参加者 4

名の日本語による対話を含むが，本稿では，日本語が母語の参加者 6 名のみで構成されるセッションに焦点を絞った．

2.2 アノテーション

教師あり機械学習の枠組みを用いるために，コーパスに対して一定の基準に従ったアノテーションを付与する必要がある．本稿では，可読性の高い音声認識出力の開発を目的としたプロジェクトである NIST Rich Transcription Project のタスクの一つである Metadata Extraction にて提案された Simple Metadata Annotation Specification [5] をもとにアノテーションスキームを定義し，非語彙的な表現と節境界の情報を付与した．

2.2.1 非語彙的表現

非語彙的な表現については，その出現範囲を示すラベルを付与した．[5] は filler types として非語彙的な表現を `filled pauses`， `discourse markers`， `explicit editing terms`，`asides / parentheticals` の 4 つに分類した．我々は更にコーパス内に多く含まれる笑い (`laughs`) を 1 つのタイプとして考案した．それぞれのタイプとその例を表 1 にまとめた．

表 1　非語彙的な表現のタイプ

タイプ名	例
`filled pauses`	「あー」「えっと」
`discourse markers`	「なんか」「なんー」
`explicit editing terms`	「つまりこう」
`asides/parentheticals`	「なんてんすか」
`laughs`	「エハハッ」「フッフ」

本稿では，表 1 に示したタイプまでは分類せず，非語彙的な表現か否かのみの情報をコーパスに付した．

2.2.2 節境界

節境界に関しては [5] で定められた 6 つのクラスにもとづいて定義し，節の境界部に付与するものとした．他クラスに分類されない一般的な発話を `statement`，相手話者への質問や自問自答などの発話を `question`，質問への返答ではない相槌を `backchannel`，「〜だったら」や「〜のとき」などの書き言葉における節と捉えられる発話の切れ目を `clausal`，割り込みや言い淀みなどによる不完全な発話を `incomplete`，「〜して，〜する」のような並列的な表現の境界を `coordination` として定義した．表 2 にそれぞれのクラスの例を示した．例内に含まれるスラッシュ記号 (/) が発話の境界である．

表 2　節境界のクラス

クラス名	例
`statement`	「それは困ったなー/」
`question`	「どんなん見るんすか/」
`backchannel`	「はーいはい/」「はいええ/」
`clausal`	「まじめなー相談をしてるとしてても/」
`incomplete`	「あの顔が/」「ちっちゃいほう」
`coordination`	「うれしくてなー / そんで…」

3　系列ラベリングによる節境界同定手法

前節で言及したアノテーションを施した ESP_C コーパスに対し，統計的機械学習の枠組みによって節境界の推定を行なう，この問題は，表層的な発話の系列から、節の境界とそうではない部分を示す隠れ系列を推定する系列ラベリングの問題として考えることができる．そこで，節境界の推定を行なうモデルを作成するために，学習器として Support Vector Machines (SVMs) [9] を適用した．2 次の多項式カーネルを用い，窓幅を前後 4 文字に設定し，推定する文字を含めて全 9 文字の情報を素性として組み込んだ．また，SVMs による推定によって得られたラベルについても前 4 文字を素性として用いた．また，対話の各参加者で長時間発話していない箇所が存在した場合，発話内容が断絶している可能性が非常に高いため，実験的に定めた 2000ms. のポーズがある箇所については確実な境界であると考え，それを跨いで素性を使用することはしなかった．

使用した素性については表 3 に示した．

表 3　使用した素性

素性	説明
文字	文字そのもの
文字クラス	かな, カナ, その他
文字の母音	a, i, u, e, o, n, - (長音), x (その他)
文字の子音	a, k, s, ..., n, - (長音), x (その他)
形態素	形態素解析器による分かち書き出力
品詞	形態素解析器による品詞大分類の推定出力
相手発話状態	発話時の相手話者の発話状態

形態素と品詞の推定には MeCab ver. 0.96 + IPA 辞書 ver. 2.7.0 [1] を用いた．音声対話の書き起こしに対しては，形態素解析器による解析誤りが多分に含まれることが予想されるため，冗長的な解析により 3-best 解を出力し，それらを全て素性として組み込んだ．相

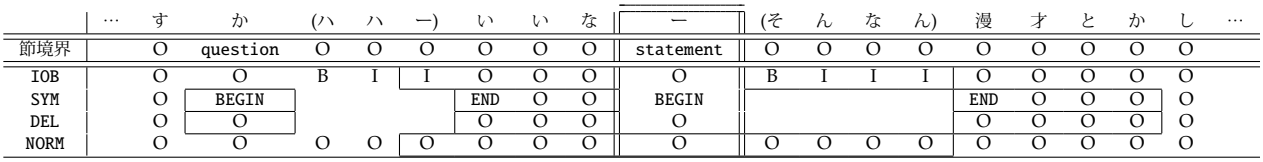

図 2　非語彙的表現の実験設定毎の素性エンコーディングの方法 (括弧内は非語彙的表現)

手発話状態は，推定を行なう時点での対話の相手の発話状態を示す情報を素性とした．これは「発話開始」,「発話中」,「発話終了」,「発話なし」,「発話引き継ぎ」の 5 クラスで分類される．

以上の情報に加え，我々は非語彙的表現の存在についても素性として組み入れることにした．

非語彙的表現の情報を表現するための方法として以下の 4 つの実験設定を考えた．非語彙的表現の出現箇所を IOB2 チャンクタグセット [7] によってエンコーディングする方法 (IOB と呼ぶ), 出現箇所をより抽象化し，出現前後に非語彙的表現の出現直前 (BEGIN)・直後 (END) であるという情報を付す方法 (SYM(ybolized)), 非語彙的な表現の出現位置を削除する方法 (DEL(eition)), そして，特に非語彙的表現に関しては何の情報も用いず表 3 の素性のみを使用した場合 (NOR(mal)) である．IOB と NOR では、非語彙的表現と考えられる範囲の文字情報も素性として組み込み，一方 SYM と DEL では非語彙的表現の文字情報を素性に含めなかった．各設定での非語彙的表現の素性エンコーディングについて図 2 にてまとめた．各行で四角で囲まれている範囲が使用した非語彙的表現の素性である．

4　評価実験

前節で述べた提案手法の性能を評価するための実験を行なった．タスクは 2 種類で，一つは節境界を同定する問題で，もう一つは節境界を同定した上でそのクラスを推定する問題である．

4.1　実験データ

実験に使用するデータは ESP_C コーパスの中の 4 セッション 120 分の書き起こしである．2 名ずつの対話で参加者は日本語が母語の話者の 5 名である．

実験データ内に非語彙的表現は 4,107 トークン，1,111 タイプ存在している．

また，節境界は 4,311 あり，クラス毎の出現数の内訳については表 5 の下部に付した．

これらのデータについて，3 節で述べた実験設定 IOB，SYM，DEL，NORM の 4 つの場合で性能比較を行なう．4 分割交差検定を行ない，評価尺度としては精度・再現率・F 値を用いるものとする．

4.2　節境界同定タスク

表 4 に各設定での節境界を同定を行なった結果を示す．

表 4　節境界同定タスクの結果

	IOB	SYM	DEL	NORM
Precision	92.7	89.9	90.1	90.3
Recall	87.0	85.0	83.0	84.8
F-measures	**89.8**	87.4	86.4	87.6

この表から，どの実験設定においても，SVMs の音声対話の節境界同定タスクへの適応は F 値で 85 を越える性能を示している．なかでも，非語彙的を IOB2 チャンクタグセットの形でコーディングされた状態で素性に組み込む IOB が最も高い性能を示した．非語彙的表現の文字情報を削除する SYM と DEL では，それら表現の文字情報を含む NORM より低い性能を示した．

以上のことから，非語彙的な表現は抽象化や削除をせずに全体の情報を用いることによって，節境界同定のタスクにおいて有効に働くことが分かった．

4.3　節境界同定・クラス分類タスク

次に問題設定を変え，2.2.2 にて述べた節境界の 6 つのクラスまで推定するタスクでの性能を調査する．

実験の結果を F 値で評価したものを表 5 に示す．これを見ると，各クラス毎に高い性能を示す実験設定が異なっていることが分かる．

節境界同定のタスクにおいて，他の実験設定と比較して低い性能を示した DEL が，このタスクでは高い性能を示している．事例数の最も多い statement クラスでは僅かながら前タスクで最も高い性能だった IOB よりも高い値を持ち，また，clausal や coodination クラスでは他と比較して劇的な高性能を示している．この原因として考えられるのは，実験設定 DEL では，非語彙的な表現を削除したので，所定の窓幅内で周辺文脈を多く見ることができたため，クラスの決定に周囲の文脈が必要なものに対し，高い性能を示したのであると考えられる．逆に，question や backchannel では周辺文脈に寄らない状態でクラスを決定しやすいために，非語彙的な表現を用いた識別で高い性能を得る

ことができたのであると考えられる．

非語彙的な表現について特別な考慮を行なっていない NORM では，他の実験設定を越える性能を示したのが，incomplete クラスのみであった．incomplete は，他話者からのインタラクションや言い淀みなどの理由で発話者が意図しない状態での発話の終了時に出現するクラスであるので，他のクラスとは異なる傾向が見られたのであると考えられる．

表 5　節境界同定・クラス分類タスクの結果 (F 値)

	IOB	SYM	DEL	NORM	事例数
statement	74.2	71.7	**74.6**	72.4	1,644
question	**78.4**	77.6	75.3	78.1	792
backchannel	**85.2**	82.7	81.4	84.0	1,080
clausal	62.3	61.3	**74.0**	61.0	425
incomplete	53.1	49.6	53.4	**53.7**	182
coordination	45.8	44.9	**50.7**	43.1	168

5　まとめと今後の課題

本稿では，統計的機械学習の枠組みを用いた，音声対話内の非語彙的な表現を手掛かりとした節境界同定の手法を提案した．

節境界の同定タスクにおいては，F 値で 86.4 から 89.8 の性能を性能を得ることができた．また，節の境界同定とクラス分類を同時に行なうタスクにおいても，クラスによっては高い性能を得ることができた．

また，本稿では従来はノイズとして処理されていたフィラーや言い淀みなどの情報が節の境界を同定する上での重要な要素となりえることを示した．非語彙的表現の出現位置について予め推定しておくことによって，節境界の性能を向上させることができた．

本手法によって，音声対話の所定の処理単位への分割が可能となり，書き起こしの可読性の向上や自然言語処理の各アプリケーションへの応用が可能となる．

今後の課題としては，本手法の堅牢性を確認するために，異なるデータセットやアノテーションスキームにおける性能を調査することがある．特に，本稿では対話の人手による書き起こしを対象に行なったが，実際のアプリケーションを考える上では，自動音声認識の出力に対して本手法を適用させることが必要である．

参考文献

[1] MeCab: Yet another part-of-speech and morphological analyzer, http://mecab.sourceforge.net/.

[2] N. Campbell. Towards conversational speech synthesis; lessons learned from the expressive speech processing project. In *Proceedings of Sixth ISCA Workshop on Speech Synthesis*, pp. 22–27, Bonn, Germany, 2007.

[3] P.-Y. Hsueh, J. D. Moore, and S. Renals. Automatic segmentation of multiparty dialogue. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 273–280, 2006.

[4] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1526–1540, 2006.

[5] NIST Speech Group. Simple metadata annotation specification version 6.2 – February 3, 2004. Technical report, Linguistic Data Consortium, 2004.

[6] G. Redeker. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics*, 14:367–381, 1990.

[7] E. Sang and J. Veenstra. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pp. 173–179, Bergen, Norway, 1999.

[8] K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara. Identification of “sentences” in spontaneous Japanese - detection and modification of clause boundaries -. In *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 183–186, 2003.

[9] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience Publication, 1998.

[10] D. Verdonik, M. Rojc, and M. Stabej. Annotating discourse markers in spontaneous speech corpora on an example for the slovenian language. *Language Resources and Evaluation*, 41(2):39–68, 2007.

[11] 柏岡．独話データのポーズ単位を利用した節境界判定．情報処理学会研究報告, 2005-SLP-57-15, 2005.

[12] 牧本, 吉川, 柏岡, キャンベル. 統計学習を用いた対話からの非語彙的表現の抽出．情報処理学会研究報告, 2008-SLP-70-30, 2008.