韻律情報を含んだマルチモーダル対話コーパスの検討*

桐山　伸也，竹林　洋一，北澤　茂良

静岡大学　情報学部

1　はじめに

GUI に代わる次世代のインタフェースとして、音声・画像を統合したマルチモーダルインタフェースが期待を集めている。音声・画像処理技術の進歩に伴い、マルチモーダル対話システムの開発が盛んであるが、ユーザの要求にふさわしい有益な情報を応答できるシステムの開発には、適切にタグ付けされた豊富な事例を含む、大規模なマルチモーダル対話コーパスが必要不可欠である。

対話システムの音声インタフェースにおいて、ユーザの発話意図を正しく理解し、分かり易い応答音声を合成するためには、音声の持つ韻律情報を適切に利用することが重要であるという観点から、我々は日本語韻律コーパスの構築を進めている。約 2,500 文からなるデータについて、音韻・韻律情報のラベル付けを行っており、現在までに、音素の境界とアクセント核の位置、並びに韻律ラベルを付与済みである。

高度な音声理解・高品質音声合成の実現のためには、大量の韻律ラベル付き音声データが必要であり、韻律ラベリングの自動化が必須技術となる。我々の韻律コーパスに付与したラベルはすべて人手によるものであり、データの信頼性は極めて高いといえる。我々は付与したラベル情報を用いて、発話の内容と韻律ラベルの対応関係の分析に基づく韻律情報の自動ラベリング手法の開発を進めている。また、マルチモーダル対話コーパスの自動インデキシングを目指し、音声ストリーム自動ラベリングの検討も行っている。

以下、2 節で構築中の韻律コーパスの仕様を概説し、3 節で韻律ラベルの概要とラベリングの自動化について述べる。さらに 4 節でマルチモーダル対話コーパスにおける韻律ラベリング手法を検討し、その応用場面を考察する。

*A Study of Multi-Modal Dialogue Corpus Including Prosody Information
Shinya Kiriyama, Yoichi Takebayashi, and Shigeyoshi Kitazawa
Faculty of Information, Shizuoka University

2　日本語 MULTEXT 韻律コーパス

2.1　仕様

EUROM1[1]と呼ばれる EU 加盟国 11 言語を対象としたデータベース作成プロジェクトの仕様に従い、MULTEXT（多言語韻律コーパス[2]）の日本語版を作成した[3]。

MULTEXT とは、EUROM1 内の 5 ヶ国語（英・仏・独・伊・西）について音韻・韻律ラベリングを行った韻律コーパスである。原稿のテキストは、1 つが 5〜6 文で構成される 40 の小節で成り立っている。人名・地名などは各国独自のものを用いるが、全体の文意は保存されている。日本語版も例に漏れず、文意を保った翻訳によって作成されている。

音声収録条件は EUROM1 に準拠している。話者は 20 代から 40 代の男女各 3 名の合計 6 名であり、朗読と模擬自発発話の二つの発話スタイルによって収録した。模擬自発発話の収録に際しては、小節ごとに想定した状況を指示し、役柄になりきるよう演じさせた。朗読との比較において、模擬自発発話スタイルの音声では、語・句単位の卓立がより明瞭となっている。

2.2　ラベル情報

現在までに、音韻ラベル・音声学の専門家の手によるアクセント核の位置情報、J-ToBI と呼ばれる韻律ラベリングスキームに基づく韻律ラベルといった情報のラベリングを完了している。音韻ラベル付与にあたっては、連続音声認識コンソーシアム 2001 年度版ソフトウェア[4]に含まれる音響モデルと認識エンジンを使用して音素セグメンテーションを行ったものを、4 人のラベラに手作業によって修正させた。韻律ラベルについては次節で詳しく述べる。

3　韻律ラベリングとその自動化

ToBI（Tone and Break Indices）とは、主に音声基本周波数（F_0）の変化パターンを記述するラ

ベリング体系であり、英語をはじめ多くの言語においてその有効性が確認されている。この日本語版が Japanese-ToBI（J-ToBI）[5]であり、英語の ToBI に基づいて東京方言の韻律的特徴を記述するために考案されたものである。

J-ToBI ラベルは、単語境界の情報を記す単語層、シンボルレベルで表し得る基本的な韻律事象の系列を表記するトーン層、各韻律境界における区切りの深さを表す BI 層、咳払い・言い淀み・強調など韻律に関係したその他の情報を記述する miscellaneous 層の 4 層構成になっている。図 1 にラベリングの具体例を示す。

上記の 4 層のうち、単語層・トーン層は、言語情報から規則的にラベリング可能な部分が多く、比較的容易に自動化できる。BI 層の自動ラベリングも、構文解析結果を用いることにより言語情報を利用してある程度期待できる。従って、言語情報を用いて初期ラベルを自動生成し、それをラベラが手動で修正するシステムは構築できる。このラベリング支援システムによって効率よくデータ収集・分析を進めることにより、強調や疑問形といった表層文に現れない情報の自動抽出手法の検討が可能になる。

4　マルチモーダル対話コーパス

マルチモーダル対話コーパスにおける音声ストリームの自動ラベリングを検討する。3 節で述べたラベリング支援システムは言語情報を利用しており、音声ストリームの正確な書き起こしを必要とする。実用性という点からは、環境雑音の存在する中で収録したデータを、リアルタイム処理できることが望まれる。近年音声認識技術も進歩し、大語彙連続音声認識システムの性能も向上してきているが、自然発話のディクテーション能力は未だ実用レベルとは言いがたい。従ってまずは、自然発話データに対する音声認識性能、及び F_0 パターンの抽出精度の現状レベルを把握するとともに、言語情報を用いた自動ラベリング手法の通用範囲を特定する必要がある。

具体的なアプリケーションとして我々は、マルチモーダル対話コンテンツオーサリングシステムを想定している。これは、いつでもどこでもだれでも気軽に情報発信ができることを目指したもので、収録素材からの重要場面の自動抽出が鍵となると考えている。画像ストリームにおける視線や頷きなどの情報と同じく、音声ストリームにおける韻律情報が発話意図の情報を含むため有益であり、ここに自動ラベリング手法を導入したい。将来的には、マルチモーダル情報を統合的に利用することで、精度良く重要場面の自動抽出を行う機構を実現したいと考えている。

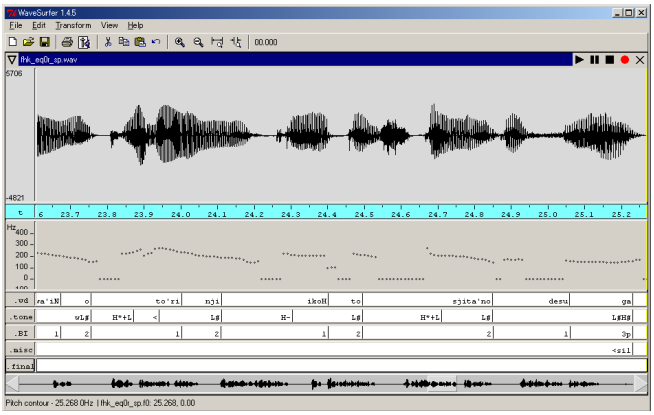

図 1:　J-ToBI ラベルの例

5　おわりに

我々が構築中の日本語 MULTEXT 韻律コーパスの概要を紹介し、韻律ラベルについてその自動化の方針を述べた。マルチモーダル対話コンテンツオーサリングシステムを想定した、韻律自動ラベリングのマルチモーダル対話コーパスへの適用を考察した。今後は韻律ラベリング支援システムの実装を行い、実データの収集・分析を通して、自動ラベリング手法の開発を進めていく。

参考文献

[1] “The SAM Projects. EUROM － A Spoken Language Resource for the EU,” ESCA, 1995.

[2] Campione, E., and Véronis, J., “A multilingual prosodic database,” proc. ICSLP98, pp. 3163-3166, 1998.

[3] Shigeyoshi Kitazawa, “Preparation of a Japanese Prosodic Database,” The ELRA Newsletter, vol.6, no.2, pp.4-6, 2001.

[4] 河原他, “連続音声認識コンソーシアム2001年度版ソフトウェアの概要,” 情処研報, 2002-SLP-43-3, 2002.

[5] J. J. Venditti, “Japanese ToBI Labeling Guidelines,” Technical Report., Ohio-State University, 1995.