2—6—7 日常会話における句末の音響・韻律的特徴の分析 *

石井カルロス寿憲 & ニック・キャンベル (JST/CREST at ATR/HIS Labs.)

1. はじめに

日本語では発話音声の句末の韻律は、疑問などの文のモダリティを表示する機能、フォーカスを表示する機能、大きな意味の区切りを示したり、発言がまだ終わっていないことを示す機能など、文法的な機能や話し手の意図や態度など、重要な役割を果たしている。

言語学や音声学の分野では文末の音調の分類を提案した研究[1,2,3]は多いが、機械による自動分類における研究は少ない。

また、朗読音声に比べ、日常会話では句末の韻律の変動が多く、日常会話に対応するための韻律ラベルとしてX-JToBI [3]が提案されているが、自動ラベリングにはまだ至っていない。

本研究は CREST/ESP の発話様式プロジェクト[4]で、韻律データベースを作成するための韻律の自動ラベリングを目的とし、句末の韻律の記述と自動分類に焦点を当て、人間の知覚によって分類化されたものと音響・韻律的特徴との対応を調べた。

2. 分析単位

データとしては、CREST/ESP プロジェクト上で録音されている自然発話音声データを用いた。発話単位としては、韻律句を扱うことにした。韻律句の区切りとしてはあきらかなポーズ、または明らかなピッチの立ち上がりが知覚される場合に半自動で行った。日本人話者一人における自然日常会話（親との会話、会社への電話）を含めた404個の韻律句を分析対象とした。

句末の定義に関しては、句末音節のV部分あるいはVN部分（つまり句末音節頭の子音を除いたもの）を“句末”と呼ぶことにした。リズムビート位置（P-Center）は音節内の母音の開始時点に近い[5]という主張から、句末を母音の開始時点から測ることにした訳である。

句末音節の切出しも音声波形のパワーと周期性の特徴を利用し、半自動で行った。

3. 句末のカテゴリー化・ラベリング

[2]では終助詞の音調の種類を次のように分けている：

1a 低く付く	例：ナ┐イネ
1b 低く付き、更に下降する	例：ナ┐イネ┐ー
2a 高く付く	例：ナ┐イ┌ネ
2b 高く付き、長く維持される	例：ナ┐イ┌ネー
2c 低く付き、上昇する	例：ナ┐イネ┌ー
3 高く付き、下降する	例：ナ┐イ┌ネ┐ー

X-JToBI では次のような句末トーンが提案されている：｜ L% (= 1a), L%+H% (= 2a,2b), L%+HL%(= 3), L%+LH%(= 2c), L%+HLH% ｜。しかし、1b は記述されてなく、句末の伸縮感覚の情報も完全に表現されていない。

本研究は以上の文献を考慮し、次のような要素を句末の分類化として提案した。

- 句末の長さ：Short (S), Long (L), Very Long (VL), Extremely Long (EL)
- 句末のトーン：Flat-Rise (FtRs), Rise (Rs), Flat (Ft), Fall (Fa), Flat-Fall (FtFa), Fall-Rise (FaRs)
- ピッチの建直しの有無：Reset
- 発声タイプ：Modal (M), Creaky (C), Breathy (B) (笑いなどによる息漏れ), Devoiced/Deleted (D) (母音の無声化、脱落), Low Energy (L)

イントネーションの研究においては F0 のみに基づいたものが多く、発声タイプの分類はあまり触れられていない課題であるが、これらの発声タイプは自然発話に頻繁に現れ、｛C,B,L｝タイプでは特に、発声区間で抽出した F0 値は M タイプに比べて信頼性が低いので、F0 抽出において注意が必要である。尚、本研究ではこのようなラベルも付与することにした。

これらのラベリング作業は母語話者1名が行い、ラベリングに疑問をもったサンプルは母語話者3名で議論して決めたものである。

4. F0 の抽出法

ここでは有声・無声判断における問題とピッチ知覚において重要な F0 値の選択に焦点を当てる。

F0 の抽出法としては、自己相関係数に基づいた手法を用いた。具体的には、音声波形に LPC 逆フィルターを通して残差波形を求め、その残差波形に LowPass フィルターを通して自己相関関数 Rxx を求める。自己相関関数を正規化したもののピークを検出し、F0 の候補とする。

従来は、正規化として $Rxx(i)/Rxx(0)$ を用いることが多いが、$Rxx(i)$ は $N-i$ 個の掛け算の和から計算され、$Rxx(0)$ は N 個の掛け算の和から計算されるため、i が大きくなるにつれて $Rxx(i)/Rxx(0)$ は小さくなる。そうなると、有声・無声の判断において、全 i に一定の閾値を決めることは適切でない。ここでは

$$\frac{N}{N-i}\frac{Rxx(i)}{Rxx(0)} \qquad (1)$$

のような正規化方法を用いることにした。このような正規化によって、自己相関関数の問題点は抑えられ、より適切な有声・無声の判断が得られる。

F0 の後処理(信頼性の低い値の削除)としては次のようなステップを提案した：

- 正規化した自己相関係数がある閾値を超えないものは削除
- 孤立点(isolated points)を削除
- マスキング効果[6]を考慮し、句末でパワーが 50ms の区間で 6 dB 以下落ちた時点の F0 値を削除

これらの制約により、人間のピッチ知覚に、より影響を与える F0 値が求められる。

* Analysis of acoustic-prosodic features of phrase finals in Spontaneous Speech, by Carlos Toshinori Ishi & Nick Campbell (JST/CREST at ATR/HIS Labs.)

5. 音響・韻律的特徴

- 句末の持続時間（*dur*）
- F0 の傾き：句末内で検出された F0 値を用いて１次回帰分析により、傾きを求める（*F0slope1*）。但し、句末が長い場合（120ms 以上）は 2 等分し、それぞれの区間で傾きを求める（*F0slope2a* と *F0slope2b*）。
- F0 の動き：句末を 2 等分し、各区間で得られた F0 のターゲット値の差分（*F0diff*）。ターゲット値は[7]で提案されたように区間の後半部分の平均値を用いている。
- F0 立て直し：句末前のターゲット値と句末の前半部分のターゲット値の差分（*F0reset*）。

6. 分析結果

ラベリングされたカテゴリー毎に測定した音響的パラメータを整理した。図１は各パラメータ（*F0slope1, F0slope2a, F0slope2b*）の各カテゴリーにおけるヒストグラムを表示している。

図から導けるように、区間全体の傾き *F0slope1* よりも、2 等分した後半の傾き *F0slope2b* の方が特に傾き 0 の周辺でよりよい弁別性を示している。*F0slope2a* の場合は弁別性が見えず、F0 の傾きは前の区間の F0 に影響されることが原因と考えられる。*F0diff* においては、*F0slope2b* と似たような傾向を示し、図は省略する。

FtRs と *Rs*、または *FtFa* と *Fa* の区別は *F0slope2b* でもはっきり見えず、これらの区別は句末の長さによって行うことにする。

7. 句末の自動分類

分析で得られた *F0slope2b* と *F0diff* の閾値を設定して句末音調の自動分類を試みた。その結果、*F0slope2b* を用いた場合 61%、そして *F0diff* を用いた場合 63%の認識率が得られた。この僅かな差はいずれのパラメータも１次回帰分析によって求められるので関連が大きいということが理由であるが、*F0diff* の方がセグメントの長さ情報も考慮していることから傾きよりも適切であると考えられる。

ラベリングにおいて、*Fa* と *Ft* の区別と、*Rs* と *Reset* + *Ft* の区別が難しかったというラベラーの感想は、自動分類で得られたこれらのカテゴリーの混乱が多い結果に反映している。これはこれらのカテゴリーが知覚的に区別しぬくいと捉えられ、カテゴリーの融合も可能と考えられる。

ピッチの立て直しにおいても、*F0reset* の閾値を設定して自動分類を行った。その結果、83%の正しい認識が得られた。

発声タイプのラベルにおいては、{C,L,D}では特に F0 抽出の問題があり、*F0slope* が計算できないケースが多く、認識率の計算から外したが、これらのケースは知覚的にはほとんど Ft か Fa、JToBI で言えば L%としてラベルされた。

8. おわりに

句末の韻律のカテゴリーを記述するため、知覚可能なピッチの動き、長さ、発声タイプを考慮して分類化してラベルしたものと、F0 の動きを定量化した音響的特徴との対応を調べた。分析の結果、カテゴリーを弁別するための閾値を求め、自動分類に応用した。F0 抽出の信頼性の低い区間、知覚的にも区別しにくいカテゴリーなどが原因で、自動分類の結果はよくなかったが、今後これらの問題点を解決する予定である。また、発声タイプの自動検出も検討している。

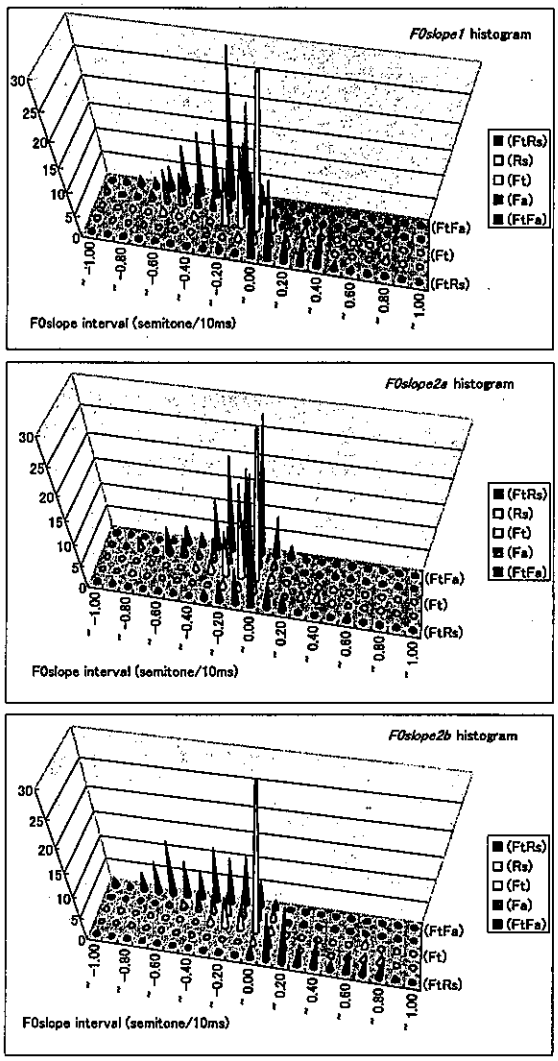

図１．各音響的パラメータのトーンのカテゴリーにおけるヒストグラム

参考文献

[1] 土岐哲「発音・聴解」外国人のための日本語例文・問題シリーズ 12，荒竹出版，37-40. (1987)

[2] 服部匡「終助詞の音調について」同志社女子大学日本語日本文学，第 14 号，1-16. (2002)

[3] 菊地、五十嵐、米山、前川「X-JToBI リファレンスマニュアル ver.1.3」11-42. (2002)

[4] The JST/CREST Expressive Speech Processing project, introductory web pages at: www.isd.atr.co.jp/esp

[5] Scott, S. "P-Centres in speech – an acoustic analysis," PhD thesis, Univ. College London. (1993)

[6] Zwicker, E. "Dependence of post-masking on masker duration and its relation to temporal effects in loudness," *JASA*, Vol. 75, Issue 1, pp. 219-223. (1984)

[7] 石井、広瀬、峯松「Investigations on a quantified representation of pitch movements in syllable units」日本音響学会春季 2002 年, Vol. I, 419-420. (2002)