多様な音声表現コーパスにおける句末音調のクラスタリング

菊池 英明（早稲田大学 人間科学学術院）†
宮島 崇浩（早稲田大学 人間科学学術院）
沈 睿（早稲田大学 人間科学学術院）

Clustering of Boundary Tones at the Accentual Phrase Edge in the Expressive Speech Corpus

KIKUCHI Hideaki (Faculty of Human Sciences, Waseda University)
MIYAJIMA Takahiro (Faculty of Human Sciences, Waseda University)
Raymond SHEN (Faculty of Human Sciences, Waseda University)

１．はじめに

表現豊かな音声が伝える様々な情報について、科学的解明や工学的応用の関心が高まっている(Erickson(2005), Schuller(2009))。発話の速さや大きさ、イントネーション、声質など、音声表現を豊かにする音響特徴は多数あるが、その中でもアクセント句末の音調が様々な非言語的情報を伝達することがわかっている。Venditti et al.(1998)は、アクセント句末に生じるピッチの変動を”BPM: Boundary Pitch Movement”と表現して、日本語東京方言における句末音調(ピッチの変動のない音調は含まない)について、生成・知覚の双方の観点で 5 種類の音調が独立して存在することを明らかにした。日本語話し言葉コーパス(CSJ: Corpus of Spontaneous Japanese)(CSJ(2011))には X-JToBI のスキーム(前川ら(2001))に基づいてラベリングがなされており、付与されたラベル系列のパタンからは、日本語(の主に東京方言)の話し言葉においては主に 7 種類の句末音調(ピッチの変動のない音調を含む)が存在するといえる(前川(2011))。岩田ら(2012)は、対話調の演技音声資料の文末音節の F0 形状をクラスタリングし、言語学分野における分類と対応させながら、代表的な 6 種類を選定した。

筆者らは、表現豊かな音声の特性を調べることを目的に、声優や俳優などに多様な状況設定を与えて演技音声を収集することにより「表現豊かな音声コーパス」を構築している(菊池ら(2012b))。このコーパスには、同一の発話内容に対して多様な表現で発声された音声が多数収録されているため、話者や内容を統制した条件で句末音調の変動を分析するのに適している。菊池ら(2012a)では、句末モーラにおける F0 変動のパタンを観察して、多様な音声表現に伴う多様なパタンがあらわれていることを確認した。本稿では岩田ら(2012)と同様のクラスタリング手法を用いて表現豊かな音声の F0 変動のパタンを自動分類し、形状の類似性に基づいた分類がどのようになるかを調べた結果を報告する。

２．表現豊かな音声コーパス

筆者らは、声優や俳優に指示を与えて多様な音声表現を収集してコーパス(通称「千の声コーパス」、以降”SEN”の略称を用いる)を構築する試みを 2008 年より続けている。指示の具体的な例を表 1 に示す。以下では、こうした指示を受けて 1 名の 40 代女性声優が発声した発話内容「あーそうですか」の 100 発話のデータを用いる。Miyajima et al.(2011)はこれらのデータについて、「怒り」「喜び」「幸福」などの基本感情語を指示して演技者に表現を委ねる従来的な収集方法によって得られたデータとの比較を行い、物理的・心理的に多

† kikuchi@waseda.jp

様性が高いことを報告している。SEN の収集方法の詳細や多様性の検証については Miyajima et al. (2011)を参照されたい。

なお、この 100 発話には分節単位ラベルと X-JToBI ラベルを付与しており、以降の分析ではこれらのラベルを用いる。

表 1 表現豊かな音声表現を得るための指示の例

共通	発話時の場所・状況	大家族を取り扱った特集において(テレビ番組)
	発話者と聞き手の関係	親子
聞き手	年齢／性別	１０歳未満／男
	職業・役柄	小学生
	人物像	典型的なやんちゃな小学生。元気があり待っている状態
発話者	年齢／性別	３０代／女
	職業・役柄	主婦
	人物像	元ヤンのヤンママと言った感じ。言葉遣いはキレイではない。
	発声時の背景	子供のだらしなさに対し、思わず声を張って叱る様子

３．分析方法

発話末のモーラ「カ」における F0 変動のパタンをクラスタリングする。まず、F0 についてはセミトーンで話者正規化したものを 3 次の最小二乗曲線で近似する。これを始端・終端を含めた 10 点でサンプリングし、差分値を 9 次元の特徴ベクトルとしてクラスタリングした。クラスタリング方法としては Ward 法、距離の測度としてユークリッド距離を用いた。なお計算には R を利用した。

図 1 にセミトーンで話者正規化した F0 値(a)と、近似曲線(b)と、サンプリングした 10 点(c)を示す。このように目視で全ての発話について近似の妥当性を確認したところ、大きく外れたものはごく数例だけであった。無声化により F0 値が抽出できないケースや極端に短いために近似ができなかったケースを除き以下では 88 発話を分析の対象とした。図 2 に全発話の近似曲線を示す。

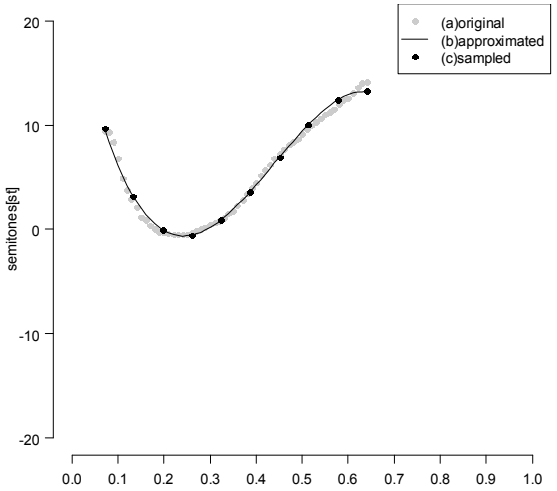

図 1　話者正規化した F0 値と近似曲線とサンプル(例)

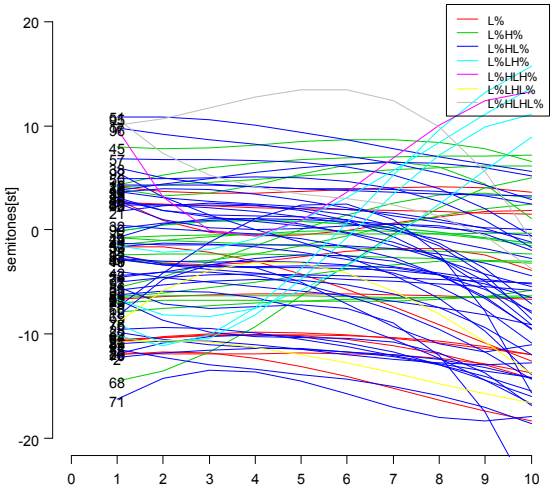

図 2　SEN の句末モーラの F0 近似曲線

４．クラスタリング結果

クラスタリング結果を図３に示す。観察しやすいように便宜上大きく 5 クラスタを認定し、それぞれにクラスタ 1〜5 と番号を与える。以下ではクラスタごとに近似曲線をプロットしてそれぞれの性質を観察する。図４にクラスタごとの近似曲線の分布を示す。

図４より、クラスタリングによって概ね F0 の形状が分離できていることがわかる。ただし、クラスタ 2 と 3 にはそれぞれ明らかに異なる形状が混在しており、下位分類(クラスタ 2A, 2B など)によって分離されている。

次に、このクラスタリング結果に基づいて元の F0 変動パタンをクラスタごとに分けて表示したものを図５に示す。これを、図６のように、人手によって付与された X-JToBI ラベルに基づいて句末境界音調(BPM: Boundary Pitch Movement)毎に観察することにより、クラスタリング結果と BPM の分類との対応関係を考察した。

岩田ら(2012)は上昇音調として疑問型上昇調と強調型上昇調を分けて扱ったが、形状を見る限り、これに相当するのがそれぞれ”LH%”と”H%”であると考えられる。”LH%”はクラスタ 4 とほぼ一対一の関係にあり、クラスタリングによってほぼ分離できているといえる。”H%”については、16 発話中 13 発話がクラスタ 2 の下位分類 2A,2B,2C に分類されている。特に下位分類 2A の 6 発話は全て”H%”の BPM が認定されており、”H%”と対応したクラスタといえる。”HL%”はいわゆる上昇下降調に相当するが、図６を見てもわかるとおり、ここにはゆるやかな下降が長いタイプと短いタイプが存在し、それぞれがクラスタ 3 とクラスタ 5 に分類されている。BPM の認定そのものにも検討が必要であるが、聴取印象の違いを調べたうえで”HL%”の下位分類の検討の必要性を示唆するものと考える。

なおその他の音調(図６の”others”)については数が少なく充分な考察ができない。今後、表現豊かな音声コーパスの資料を利用して出現頻度の少ない音調についても調査する必要がある。

５．まとめ

表現豊かな音声コーパスの一部を用いて、クラスタリング手法によって句末音調の F0 形状に基づく自動分類を行った。X-JToBI ラベルに基づく BPM の分類との対応関係を調べたところ、”LH%”と”H%”などの、クラスタと BPM との対応がよくとれる音調と、”HL%”などの、対応がとれていない音調が存在することがわかった。現在のところ、クラスタリングの特徴量として長さや高さの情報を用いていないなど、クラスタリングの精度を向上させる余地がある。また、今回は一話者の音声のみを対象としたが、ある程度の多様性は確認されているものの表現の種類には話者固有性があると考えられるため、複数話者の音声についても検討する必要がある。今後は表現豊かな音声コーパスの他のデータを用いてさらに大規模な検討を進めていく。

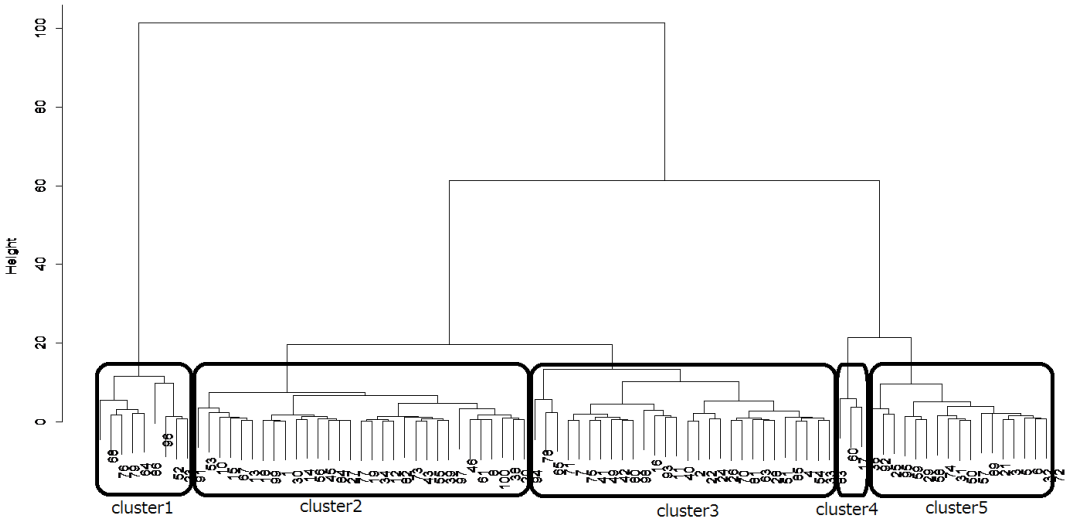

図３　クラスタリング結果
（リーフの番号は発話番号）

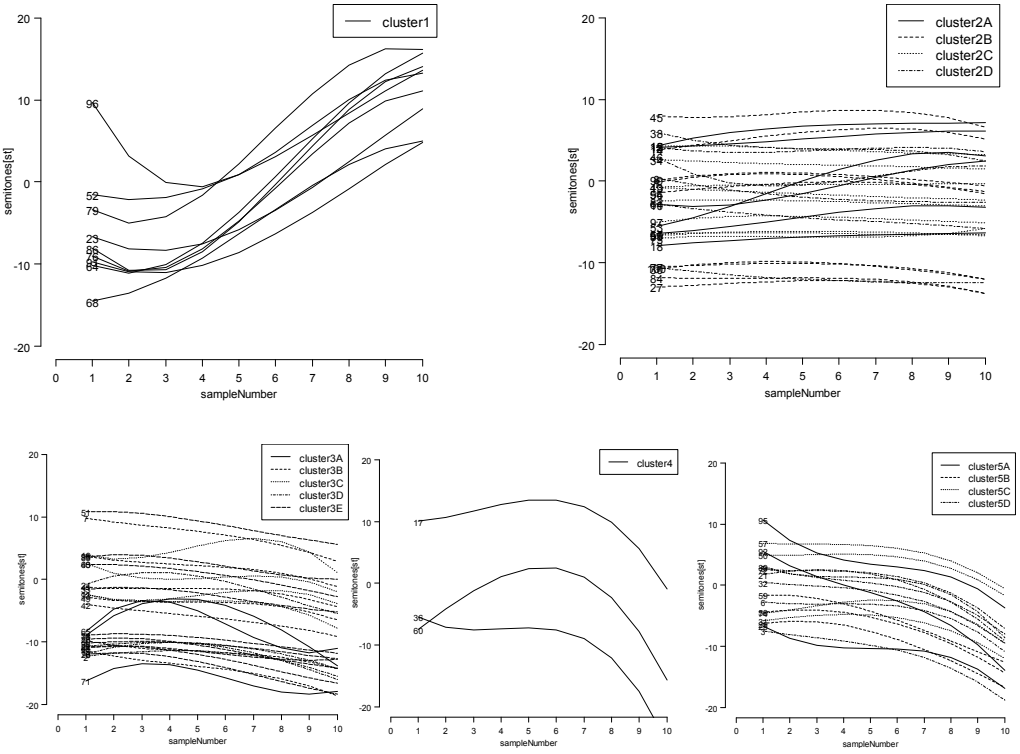

図４　クラスタごとの近似曲線の形状分布
（上段左からクラスタ 1, 2、下段がクラスタ 3,4,5。各曲線左端は発話番号。）

文　献

D. Erickson (2005). “Expressive speech: Production, Perception and Application to Speech Synthesis”, Acoust. Sci. & Tech., vol.4, no.26, pp.317-325.

B. Schuller, S. Steidl, A. Batliner (2009). “The INTERSPEECH 2009 Emotion Challenge”, Proc. of INTERSPEECH 2009, pp.312-315.

J. Venditti, K. Maeda, and J. P. H. van Santen (1998). “Modeling Japanese boundary pitch movements for speech synthesis.” Proc. of the 3rd ESCA Workshop on Speech Synthesis.

前川喜久雄, 菊池英明, 五十嵐陽介 (2001). 「X-JToBI: 自発音声の韻律ラベリングスキーム」, 電子情報通信学会技術報告(NLC2001-71, SP2001-106), pp.25-30.

前川喜久雄 (2011). 「コーパスを利用した自発音声の研究」, 東京工業大学大学院博士論文.

CSJ(2011). 「日本語話し言葉コーパス」, 国立国語研究所, http://www.ninjal.ac.jp/csj/

T. Miyajima, H. Kikuchi, K. Shirai (2011). "Collection and analysis of emotional speech focused on the psychological and acoustical diversity", Proc. of ICPhS2011, pp.1394-1397.

菊池英明, 宮島崇浩 (2012a), 「日本語話し言葉コーパスにおける句末音調のバリエーション」, 第 2 回コーパス日本語学ワークショップ, pp.351-354.

菊池英明, 宮島崇浩, 前川喜久雄 (2012b), 「表現豊かな音声の収集における多様性の追求」, 日本音響学会秋季研究発表会講演論文集, Vol.1-2-16, pp.263-264.

岩田和彦, 小林哲則 (2012), 「終助詞とその音調とによって聞き手に伝わる発話意図の分析」, 電子情報通信学会技術報告, SP2012-77, pp.31-36.

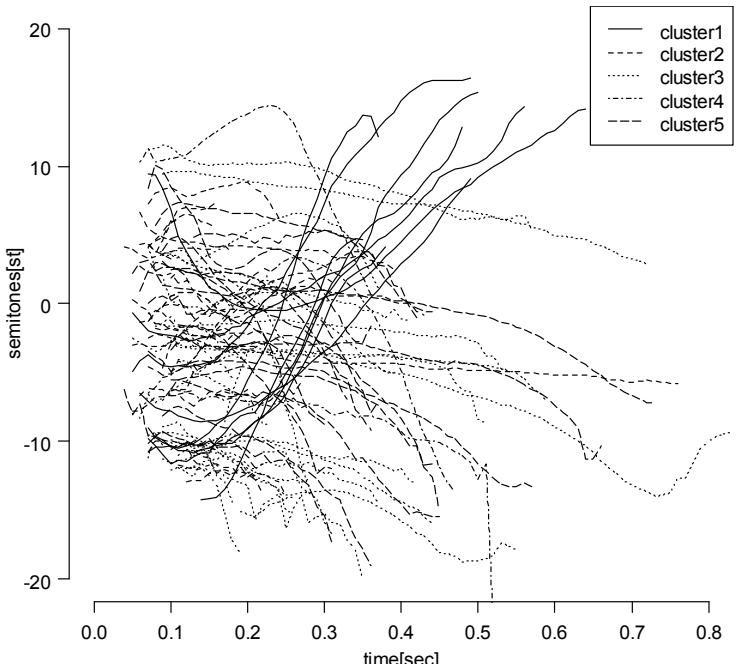

図 5　SEN の句末モーラの F0 変動

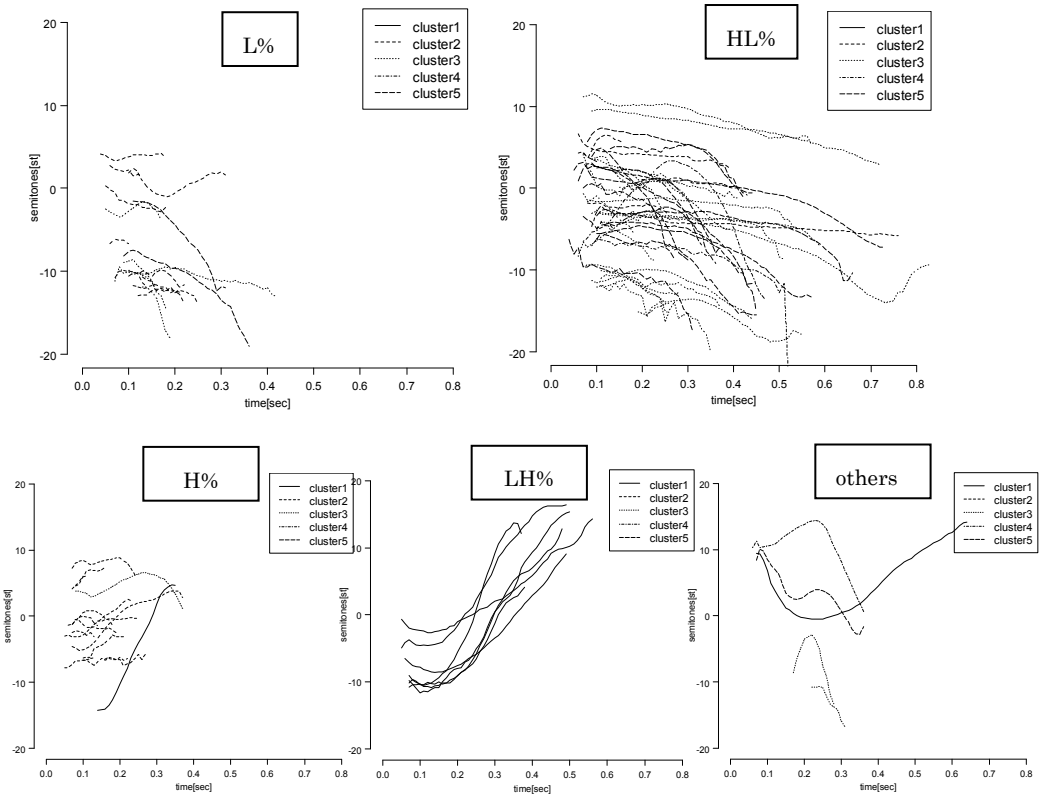

図６　人手で分類した BPM とクラスタリング結果との対応