日本語話し言葉コーパスのF0値再抽出に関する検討

著者	石本 祐一, 河原 英紀
雑誌名	言語資源活用ワークショップ発表論文集
巻	2
ページ	297-303
発行年	2017
URL	http://doi.org/10.15084/00001531

日本語話し言葉コーパスの F0 値再抽出に関する検討

石本 祐一 (国立国語研究所コーパス開発センター) *
河原 英紀 (和歌山大学)

A Study on the Revision of Annotated F0 for the Corpus of Spontaneous Japanese

Yuichi Ishimoto (National Institute for Japanese Language and Linguistics)
Hideki Kawahara (Wakayama University)

要旨

本稿では、日本語話し言葉コーパス (CSJ) の音声から高精度な基本周波数 (F0) 抽出法による F0 推定を行なった結果について報告する。CSJ に現在付与されている F0 値は、波形振幅の正規化相互相関に基づいた手法により推定されたものである。波形振幅の相関に基づく手法は波形の周期性を利用するため、周期性が乱れた箇所では本来の F0 とはかけ離れた値を誤検出したり、非周期区間とみなされて F0 の抽出が全くできないことがある。実際に CSJ に付与された F0 情報を確認すると、明らかな抽出誤りの箇所が見受けられる。F0 は CSJ に付与されている韻律ラベリングの基となっているため、F0 抽出精度は韻律ラベリングの精度にも関わる。そこで最新の高精度な F0 推定法によって F0 値を求め、現在付与されている値との違いを調べるとともに、韻律ラベルへの影響について述べる。

1. はじめに

音声において基本周波数 (F0) は重要な音響特徴量であり、音声分析やその応用に広く用いられている。そのため、音声コーパスを各種分野に供与することを考えると、コーパス構築段階で F0 値やそれに関する韻律ラベルが付与されていることが望ましい。国立国語研究所が配布している大規模な話し言葉データベースである『日本語話し言葉コーパス (CSJ)』(Maekawa et al. 2000) においては、各発話に F0 情報と日本語の韻律ラベリングスキーム X-JToBI(Maekawa et al. 2002) に基づいた Tone ラベルや Break Indices 等が付与されており、これらは話し言葉の韻律研究の進展に大いに貢献している。

しかし、CSJ に付与されている F0 をつぶさに見ると、ほとんどの区間では正しい F0 を指し示していると思われるものの、明らかな F0 抽出誤りとみられる値も散見される。本稿では、CSJ の F0 情報の改善を目指し、CSJ に現在付与されている F0 と高精度の F0 推定法によって抽出した F0 の比較を行った結果について報告する。

* yishi@ninjal.ac.jp

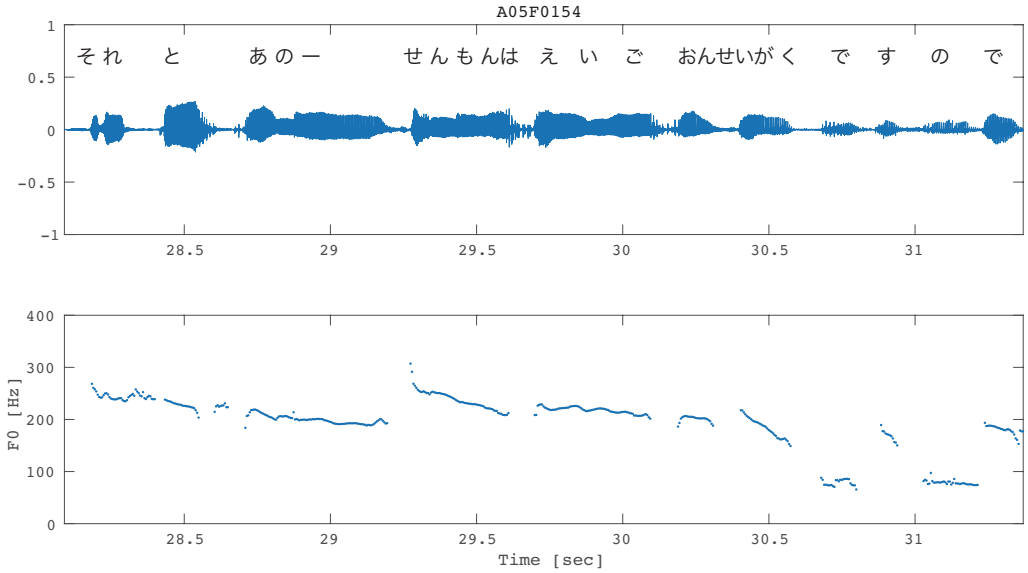

図 1　CSJ に付与された F0 情報の例（上段：音声波形、下段：F0）

2. F0 抽出

2.1 CSJ に付与されている F0

CSJ には、音声分析プログラム群 ESPS toolkit に含まれている get_f0 関数によって推定された F0 値が、5 ms 間隔で付与されている。get_f0 は波形の正規化相互相関による F0 候補算出と動的計画法によるトラッキングを用いた F0 推定法 (Talkin 1995) であり、安定して動作することから音声分析ソフトウェア Wavesurfer 等の様々なアプリケーションで同種の推定法が利用されている。CSJ に対する F0 推定結果を目視により確認すると、この手法によりほとんどの区間で正しい F0 が推定できているように思われる。しかし、中には明らかに誤った値を推定していたり、聴感上は有声と思われる区間にもかかわらず F0 が抽出できていない箇所もある。

例として図 1 に、CSJ に収録されている学会講演の発話（ファイル ID：A05F0154）の一部と、その推定 F0 を示す。女性の音声であるため 200–300 Hz 付近に F0 が現れており、「それとあのー専門は英語音声学」まではおおよそ聴感的にも疑問のない値となっている。しかし、つづく「ですので」の「で」と「の」の区間においては、それまでの F0 の流れとは大きく異なる 100 Hz 以下の低い値が推定されていることが見てとれる。この低い値は、本来の F0 の半分にあたる数値が誤って推定されていると考えられる。波形振幅の相関を利用する F0 推定法は原理的にこのような $\frac{1}{n}$F0 を誤推定しやすい傾向にあり、図 1 に例示した箇所以外でも同様の誤りが数多く生じている可能性が高い。音声コーパスにおいては F0 の概形は Tone 層のラベリングに、F0 の高さは BI 層のラベリングに関わるため、コーパスの韻律ラベルに対する信頼性向上のためにはより高精度な F0 推定が求められる。

本稿では以後、CSJ に付与されている F0 を CSJ-F0 と呼ぶこととする。

表 1　対象談話

TalkID	性別	全発話長 [sec]	有声区間長 [sec]	(F0 点数)	F0Uncertain 箇所
A05F0154	女性	829.072	444.080	(88816)	1145
A11M0469	男性	1685.598	912.135	(182427)	1016
A05F0039	女性	830.089	521.045	(104209)	956
A03M0045	男性	777.552	387.495	(77499)	905
A01F0132	女性	586.167	424.350	(84870)	163
A01F0145	女性	470.420	281.915	(56383)	157
A01M0110	男性	392.054	341.170	(68234)	135
A01M0147	男性	597.417	501.370	(100274)	128

2.2　高精度な F0 推定

著者の一人が中心となって、これまでに多数の F0 抽出手法が開発されている。ここでは、高精度な手法として最近開発された、残差と瞬時周波数に基づく基本波成分候補の抽出と Kalman filter により推定された潜在変数の軌跡による候補選択とを組み合わせた F0 推定法 (Kawahara et al. 2016, 2017) を用いる。この手法では F0 だけではなく F0 軌跡の信頼区間も得られるが、本稿では最良の候補として推定された F0 値だけを取り扱う。

なお、本手法のプログラムは本稿執筆時にはまだ非公開であるが、2017 年 8 月末を目処に公開される予定である。

以後、この手法で推定された F0 を New-F0 と呼ぶこととする。

3.　CSJ-F0 と New-F0 の差の分析

CSJ に付与されている F0 情報 (CSJ-F0) と新たに高精度な推定法で推定した F0(New-F0) の違いを調べる。

3.1　対象データ

CSJ では、Tone ラベルを付与する位置の F0 が不明確である（推定誤りの可能性がある）ことが確認された箇所に、F0Uncertain ラベルを付与している (国立国語研究所 2006)。そこで、F0 推定誤りが多い談話と少ない談話に対する New-F0 の効果の調査を目的として、CSJ に収録されている学会講演の中から F0Uncertain 箇所が多い講演と少ない講演をそれぞれ男女 2 名分ずつ選んだ。表 1 に選択された談話の情報を示す。ここで、有声区間とは CSJ-F0 が存在する区間を意味する。New-F0 では有声/無声判定を実施していないため、無声区間を含む全ての発話区間で何らかの値が得られているが、CSJ-F0 との比較のために CSJ-F0 の有声区間を New-F0 でも有声区間とみなして適用することとした。全発話長と有声区間長の比率を考えると、F0Uncertain 箇所が少ない講演に比べて、F0Uncertain 箇所が多い講演は有声区間長が短い傾向にあるが、これは CSJ-F0 において有声にも関わらず F0 が抽出できない区間が多く存在することを示唆している。ただし、前述のように New-F0 では現在のところ有声/無声判定

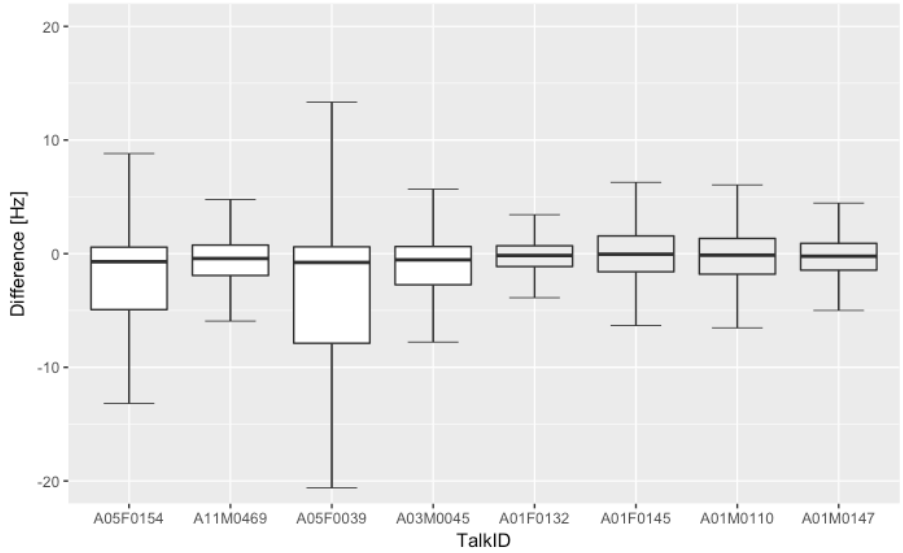

図 2　全発話における F0 差の分布

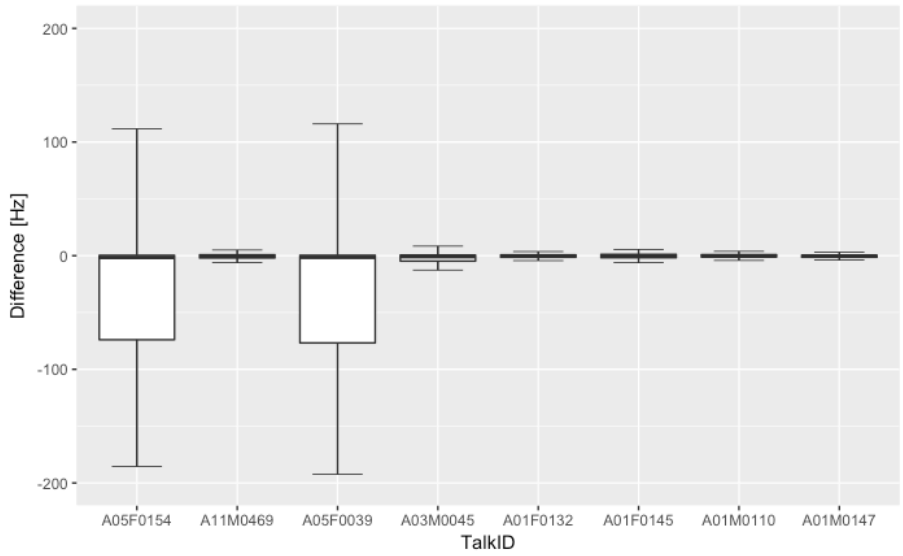

図 3　F0Uncertain 箇所を含むアクセント句における F0 差の分布

を行なっていないため、CSJ-F0 で F0 抽出されていない区間については本稿では考慮しない。

また、New-F0 はサンプリング間隔で F0 を抽出することができるが、CSJ-F0 が 5 ms 毎に付与されていることから、New-F0 でも CSJ-F0 と同時刻の 5 ms 間隔の値を抽出した。

さらに、各時刻で CSJ-F0 から New-F0 を引いた値を推定 F0 差として求めた。

3.2　結果

図 2 に談話毎の CSJ-F0 と New-F0 の差を示す。なお、図の見やすさを考慮し、ヒゲの範囲から外れた値は図中に描画していない。図の左半分が F0Uncertain 箇所が多い談話、右半分が少ない談話である。中央値に着目すると、どの談話も 0 Hz 付近になっていることがわか

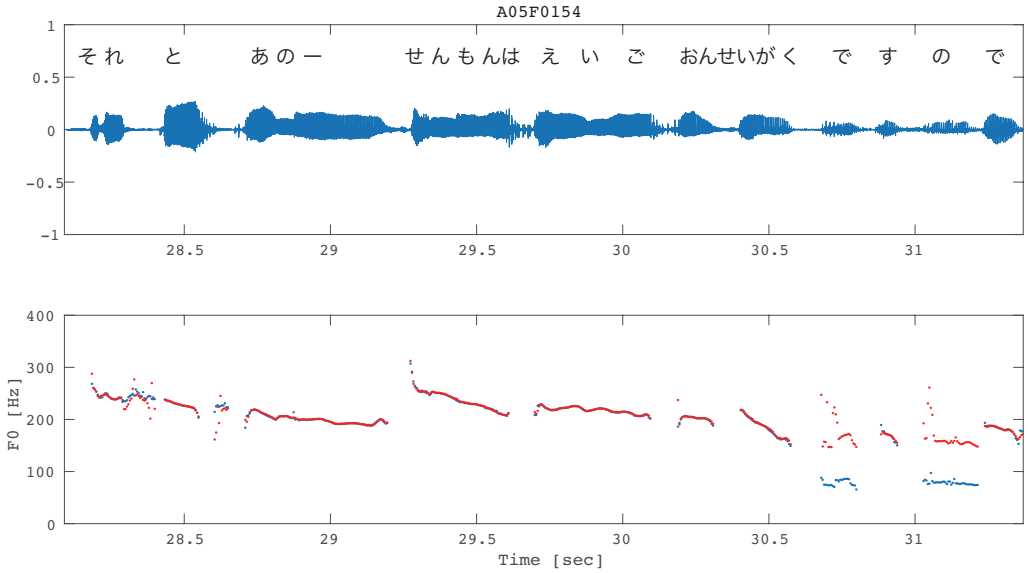

図 4　女性話者における CSJ-F0 と New-CSJ の例（下段青：CSJ-F0、下段赤：New-F0）

る。特に、F0Uncertain 箇所が少ない談話では男女の区別なくほとんどが数 Hz 以内に収まっており、CSJ-F0 と New-F0 でほぼ同じ値が推定されていると考えられる。数 Hz の違いに対し「どちらの F0 がより正確か」といった議論も成り立つであろうが、本稿で論ずることは避ける。

F0Uncertain 箇所が多い談話では、男性話者に関しては F0Uncertain が少ない談話とさほど違いはないものの、女性話者については大きくばらついている。分布は差が負になる方に広がっていることから、CSJ-F0 よりも New-F0 が高い値になっている区間が多く現れていることがわかる。また、女性話者であっても F0Uncertain 箇所が少ない談話ではこのようなばらつきは見られないことから、F0Uncertain 箇所、すなわち CSJ-F0 が不明瞭と判定されている箇所が影響していると考えられる。

そこで、F0Uncertain 箇所を含むアクセント句に対象を絞った場合について、談話毎の CSJ-F0 と New-F0 の差の分布を図 3 に示す。図 2 と図 3 では縦軸のスケールが異なることに注意されたい。図 2 でばらつきが大きかった談話については、図 3 でさらに顕著になっており、F0Uncertain 箇所で CSJ-F0 と New-F0 でまったく異なる値が推定されていることがわかる。その差が 100 Hz 近いものも多くあり、CSJ-F0 で誤抽出された区間の多くで New-F0 が正しい値を示していると考えられる。

図 4 に女性話者における CSJ-F0 と New-F0 の例を示す。なお、図 4 は図 1 と同一の区間である。まず「それとあのー専門は英語音声学」までの CSJ-F0 と New-F0 を見ると、ほとんど同じ値になっていることがわかる。このように、CSJ-F0 においてそれらしい F0 値である区間では New-F0 でもほぼ同様の値となり、F0 概形もほぼ同一となっている。一方、F0Uncertain 箇所を含むアクセント句「ですので」では、前述の通り「で」と「の」において CSJ-F0 が前後の F0 の流れよりも低い値となっているが、New-F0 は前後とほぼ同じ範囲であり、CSJ-F0

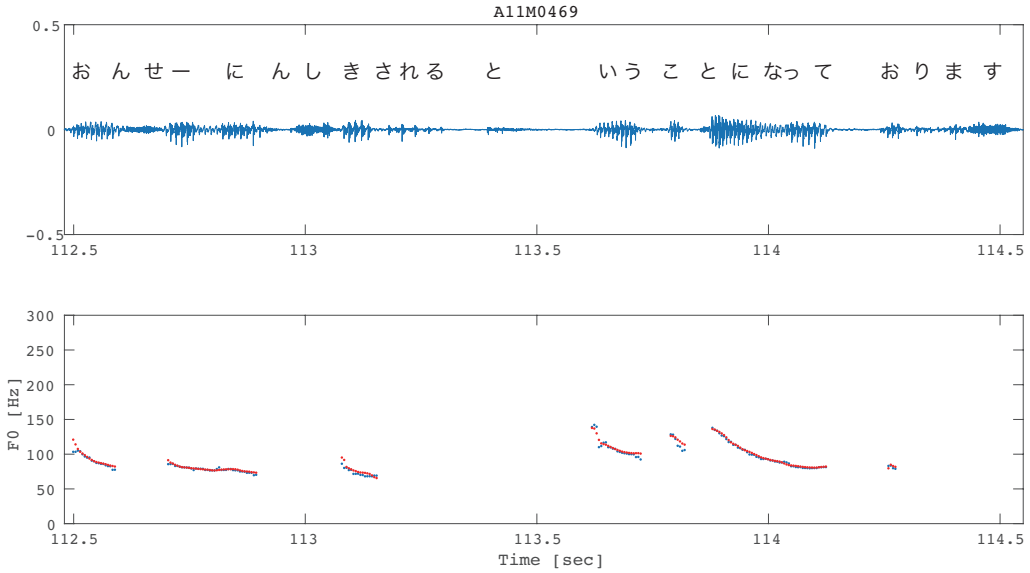

図 5　男性話者における CSJ-F0 と New-CSJ の例（下段青：CSJ-F0、下段赤：New-F0）

のほぼ 2 倍の値を示している。実際の音声を聴取すると「で」や「の」の区間の F0 が特に低下していないことが明確であることから、New-F0 で推定された値が正しい F0 と考えられる。

では何故、男性話者に対しては女性話者ほど CSJ-F0 と New-F0 の差が現れないのだろうか。図 5 に男性話者における CSJ-F0 と New-F0 の例を示す。CSJ-F0 と New-F0 はほぼ同じ値となっておりほとんど差が見られない。F0Uncertain 箇所はアクセント句「されると」と「おります」であるが、無声化や周期性の乱れによって CSJ-F0 においては周期性が認められずに F0 抽出ができていない。そのため、CSJ-F0 の有声区間に合わせた New-F0 でも F0 が存在しないことになり、本稿の基準では CSJ-F0 との違いとして現れない状態になっていた。このように男性話者では F0 の誤抽出ではなく、F0 が抽出できるかどうかが問題となっていると考えられる。CSJ-F0 とは異なる有声/無声判定を導入することで、New-F0 によりこれらの区間でも正しい F0 が抽出できる可能性があるが、今後の課題としたい。

最後に、CSJ にすでに付与されている韻律ラベルに New-F0 が与える影響について考察する。図 4 や図 5 に見られるように、CSJ-F0 と New-F0 では F0 概形にはほとんど違いがない。そのため、アクセント位置やアクセント句内の F0 の上昇や下降を表す Tone ラベルには影響しないと思われる。一方、BI ラベルはアクセント句間の F0 の違いによって判別されるため、図 4 のように F0 の高さが修正されることで、異なる BI ラベルが選択される可能性がある。BI 層への影響は今後詳細に調査する必要があろう。

4. おわりに

CSJ に現在付与されている F0 情報のうち明白な誤りであるものについて修正することを目的として、新しく開発された高精度な F0 抽出手法により F0 推定を行い、CSJ に付与されている F0 との違いを調べた。その結果、女性話者音声に対して多い傾向にある $\frac{1}{2}$F0 を誤抽出し

ている問題に対し、新手法が対応できる可能性が示唆された。今後は、男性話者音声へも対応範囲を広げ、改善した F0 情報を CSJ 利用者へ提供できるよう進めていく予定である。

謝　辞

本研究の一部は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」（2016–2021 年度）の成果である。

文　献

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara (2000). “Spontaneous speech corpus of Japanese.” *Proc. LREC2000*, pp. 947–952.

Kikuo Maekawa, Hideaki Kikuchi, Yosuke Igarashi, and Jennifer Venditti (2002). “X-JToBI: An extended J_ToBI for spontaneous speech.” *Proc. ICSLP2002*, pp. 1545–1548.

David Talkin (1995). “A Robust Algorithm for Pitch Tracking (RAPT).” Willem Bastiaan Kleijn, and Kuldip K. Paliwal (Eds.), *Speech Coding and Synthesis.*: Elsevier Science B.V.. pp. 495–518.

Hideki Kawahara, Yannis Agiomyrgiannakis, and Heiga Zen (2016). “Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis.” *9th ISCA Speech Synthesis Workshop*, pp. 221–228. (arXiv preprint arXiv:1605.07809)

Hideki Kawahara, Ken-Ichi Sakakibara, Hideki Banno, Masanori Morise, and Tomoki Toda (2017). “A modulation property of time-frequency derivatives of filterd phase and its application to aperiodicity and f_o estimation.” *Proc. Interspeech 2017*. (Accepted)

国立国語研究所 (2006).「日本語話し言葉コーパスの構築法」　国立国語研究所報告:124.