国立国語研究所学術情報リポジトリ

# 日常会話音声に対する基本周波数推定の課題


| | |
|---|---|
| 著者 | 石本 祐一 |
| 雑誌名 | 言語資源活用ワークショップ発表論文集 |
| 巻 | 4 |
| ページ | 381-391 |
| 発行年 | 2019 |
| URL | http://doi.org/10.15084/00002590 |

# 日常会話音声に対する基本周波数推定の課題

石本 祐一 (国立国語研究所コーパス開発センター) *

# Practical Issues of Fundamental Frequency Estimation for Everyday Conversation Speech

Yuichi Ishimoto (National Institute for Japanese Language and Linguistics)

**要旨**

音声コーパス構築において韻律ラベリングを行うためには、音声波形の基本周波数を抽出することで声の高さを数値化し、上昇・下降の程度を観察することが必要となる。一般に、『日本語話し言葉コーパス』に収録されているような雑音がほとんど存在しないクリーンな音声に対する基本周波数推定は容易であるが、日常場面のような周囲に様々な音が存在する環境で収録された音声に対しては各種の雑音の影響や発話の重複により誤った推定がなされる場合がある。本稿では『日本語日常会話コーパス』モニター版を基に、推定エラーが生じやすい日常会話音声に対して雑音抑圧や音源分離といった音声信号処理を利用することで、音声コーパス構築に向けてどの程度の基本周波数推定を行うことができるかを示す。

## 1. はじめに

日常生活において会話がなされるとき、周囲には会話音声以外にその環境に応じた様々な音が存在している。そのため、日常会話音声の録音を行うと、目的の音声に加えて周囲の雑音が混入することがあり、時には雑音が目的音声の聴取を妨げることが生じる。また、発話内容を聴き取ることができるため聴感上は雑音の影響が小さく感じられる音声であっても、基本周波数 (F0) の抽出を試みると日常場面の雑音の影響で正しい F0 値を得ることができない場合もある。このように、自然な日常会話に対して実験室と同等のクリーンな音声を得ることは難しく、特に音響特徴量の抽出に困難を伴うことが多い。

現在、国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」により、多様な種類の日常会話をバランス良く収録した大規模な日常会話コーパスとして『日本語日常会話コーパス』 (Corpus of Everyday Japanese Conversation, CEJC) の構築が進められている (Koiso et al. 2018)。このコーパスでは可能な限り雑音が入りにくいように収録方法を考慮しているが、やはり収録された音声の一部に前述のような問題が生じている。本稿では、2018 年 12 月から公開されている CEJC モニター版を基に収録音声のうち F0 推定の誤りが生じやすい音声を取り上げ、どのような問題があるのかを示す。また、雑音抑圧や音源分離といった音声信号処理を利用することで F0 推定のために目的音声を改善する方法について紹介し、日常会話コーパス構築への貢献の可能性を示す。

* yishi@ninjal.ac.jp

## 2. 日常会話音声の基本周波数推定

『日本語話し言葉コーパス』(CSJ)(Maekawa et al. 2000) ではヘッドセットマイクにより話者の口元に近い位置にマイクを置くことで、周囲の雑音がほどんど入らないクリーンな音声の収録を実現している。一方、CEJC では調査者が介在しない実際の日常場面における会話の収録を目指しているため、ヘッドセットマイクを話者に装着させることはできない。そこで、収録対象者ごとに胸元に IC レコーダを装着することでなるべく口元に近い位置にマイクが置かれるようにし、話者の音声以外の音が入らないように考慮している。その結果、明瞭に対象者の音声が録音できていることが多いものの、一部には分析に支障が生じる音質の音声も存在する。

日常会話音声の録音では大まかに次の 2 つのどちらかまたは両方が生じることにより問題となる。

1. 周囲の環境音が音声に重畳される
2. 話し相手の音声が対象者のマイクに入り込み、対象者の音声と重複する

本節では、それぞれの影響により F0 推定結果がどのようになるのか、また雑音抑圧や音源分離を前処理として音声に適用することで F0 推定結果がどのように変わるのかを述べる。

F0 推定は音声分析研究の初期から続く研究課題であり現在では F0 推定を行う様々な手法・ソフトウェアが存在するが、ここでは無料でダウンロード・使用することができる音声分析ソフトウェア Praat (Boersma and Heuven 2001) により F0 を算出することとする。なお、本稿では Praat version 6.0.50 を用いている。

### 2.1 周囲の雑音の重畳

日常場面においては大きさの大小はあるものの必ず周囲に会話以外の何らかの音が生じている。例えば、家庭内であればエアコンの動作音、屋外であらば自動車の走行音、店内であればBGM といったような会話音声に重畳してくる雑音が存在しており、収録時に除外できず録音音声に入り込んでしまうことがある。CEJC モニター版は大きな問題がある音声を採用しないようにデータ選択がなされているが、発話内容が聞き取れる程度の雑音の混入は許容されており、F0 推定がうまく行えない音声も存在する。

図 1 は、CEJC モニター版に収録されている会話のうち、会話 ID C002_008 の走行中の自動車内で発声された女性の音声の一部である。図 1 の点線の範囲内で「足がなかったらそうか」と発声しているが、上段に示す音声波形を見てわかるように発声していない区間にもある程度大きい自動車走行音が重畳しており、女性が小さめの声で発声していることもあって、発話区間が判然としない。図 1 中段は音声信号を時間-周波数平面上に展開しエネルギーの大きさを濃淡であらわしたスペクトログラムを示しており、自動車走行中の車内雑音は 1 kHz 以下に強いエネルギーを持ち時間変化がほとんどないことが見てとれる。つまり、自動車の走行雑音は音声波形で見ると不規則な変化として現れるが、周波数情報として捉えると時間的な変化がほとんどない定常雑音であるといえる。この音声に対して Praat により F0 推定を行うと、図 1 下段に示すように、主に音声の開始付近と音声の後半部分の値を得ることができていない。発

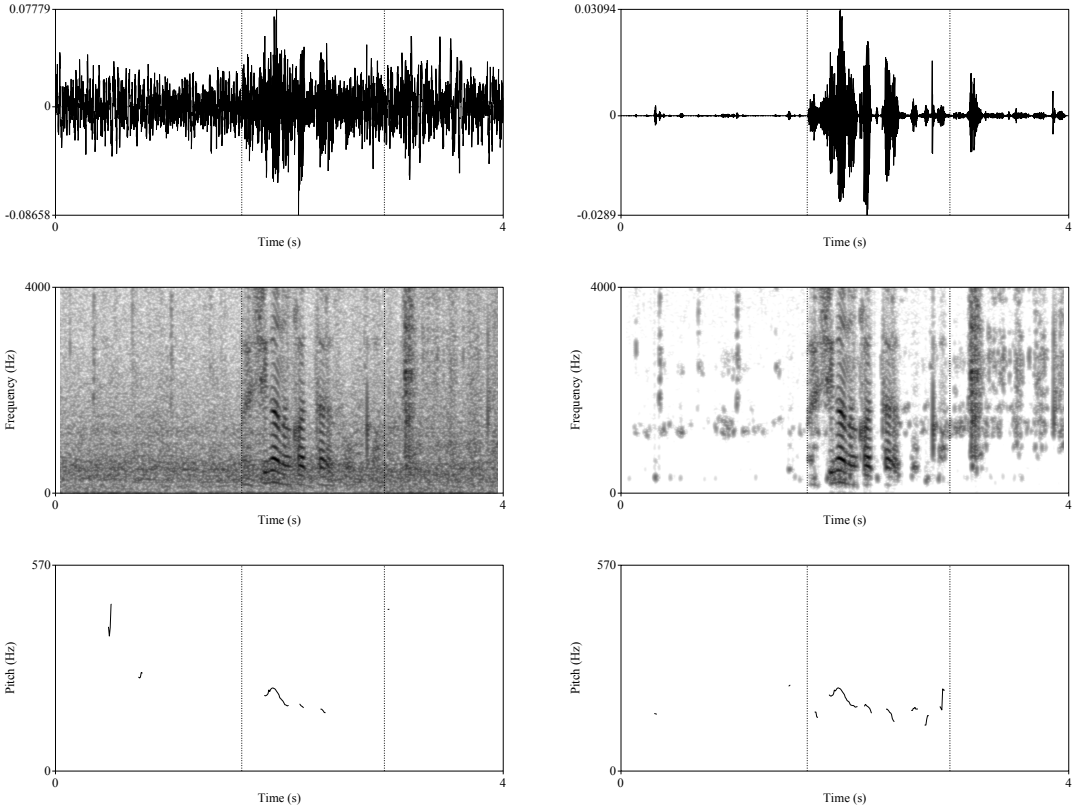



図 1　自動車走行中の会話（会話 ID C002_008）に 図 2　自動車走行中の会話（会話 ID C002_008）に
おける発話音声「足がなかったらそうか」:（上）音 おける雑音抑圧処理後の発話音声
声波形、（中）スペクトログラム、（下）推定 F0

音内容と照らし合わせると F0 が推定できなかった箇所は発話冒頭の「足」の/a/と発話最後の「そうか」に該当する。これは、自動車の走行音によって音声の振幅が小さい箇所の情報がかき消されてしまい、F0 に関わる波形の周期性を観察することが困難になったためである。

このような、雑音により F0 推定に支障が生じている音声に対して雑音抑圧処理を施すことを考える。雑音抑圧の手法も多数存在するが、本稿では Wiener filter による雑音抑圧と調波成分の再生成を組み合わせた手法 (Plapous et al. 2006) を用いることとする。この手法の MATLAB 言語による実装コードが提案者の一人によって無料で公開されているため（関連 URL 参照)、MATLAB を利用できる環境であれば誰でも容易に効果を確認することができる。

図 2 に雑音抑圧処理後の音声波形とスペクトログラム、推定 F0 を示す。図 2 上段の音声波形を見て明らかなように、自動車走行雑音が大幅に抑制されており、発話区間が明瞭に判別できるようになっている。なお、発話区間終了後に振幅の大きな波形が残っているが、これは自動車のウィンカーレバーの操作音であり、このような突発性の雑音は除去されない。音声波形と同様に、図 2 中段のスペクトログラムからも、雑音によって隠されていたスペクトル構造が現れていることがわかる。図 2 下段の推定 F0 を見ると、雑音抑圧前は推定できていなかった発話冒頭の/a/の部分や「そうか」に該当する部分の F0 値が得られている。さらに、雑音抑圧

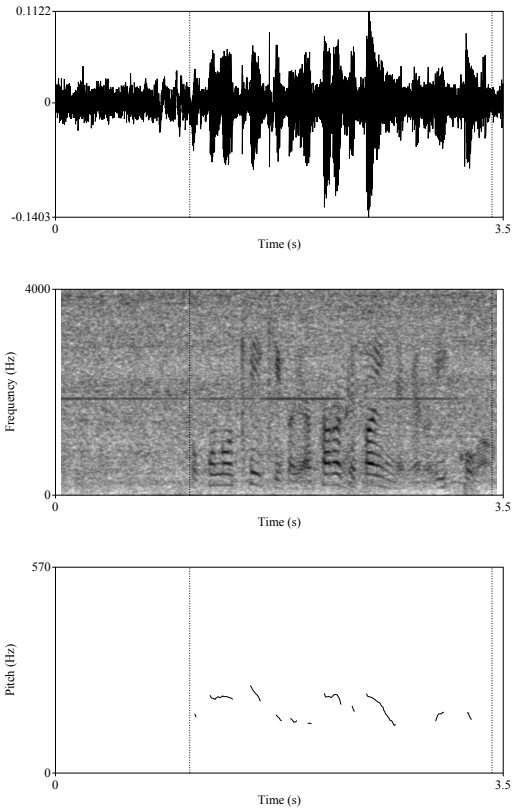



図 3　コーヒー販売店内の会話（会話 ID T009_022）における発話音声「そこのケーキがやたらでかいんだけど 一個が」：（上）音声波形、（中）スペクトログラム、（下）推定 F0

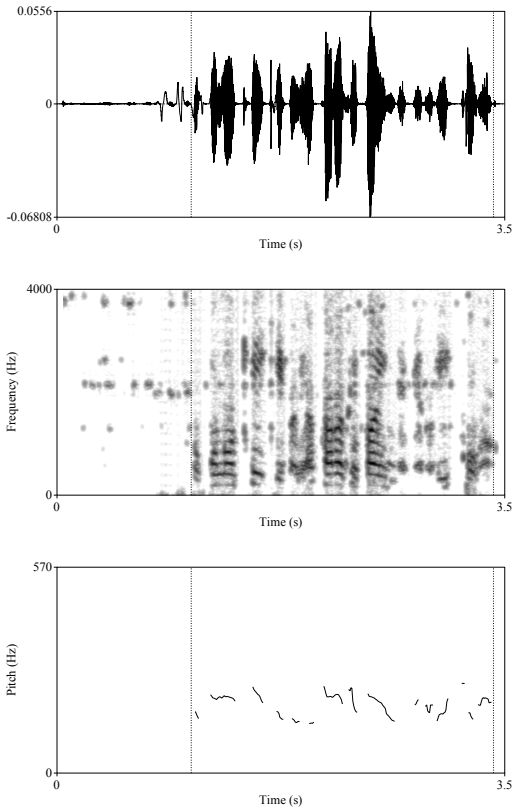



図 4　コーヒー販売店内の会話（会話 ID T009_022）における雑音抑圧処理後の発話音声

前は F0 がわずかながら得られていた発話中盤の「なかったら」の箇所も雑音抑圧後はより広い範囲で F0 が推定できていて、F0 の時間変化が観察できるようになっている。

別の例として、図 3 に会話 ID T009_022 のコーヒー販売店内で働いている女性の会話音声の一部を示す。会話 ID T009_022 では店内で稼働している機械が発する音が常に存在しており、録音音声にも入り込んでいる。図 3 上段が音声波形を表しており、点線の範囲内で「そこのケーキがやたらでかいんだけど 一個が」と発声している。雑音に関して発話区間より前の部分を見ると、音声の開始位置の判別が難しいほどの振幅で機械音が録音されていることがわかる。図 3 中段のスペクトログラムからは、機械音が特定の周波数に偏ることなく幅広い周波数帯に存在しており、先の自動車走行雑音と同様に時間的は変化がほとんど生じないことがわかる。つまり、定常雑音という点でこの店内の機械音と自動車走行雑音に大きく異なる特徴はないといえる。この機械音を含む音声から Praat で F0 を推定した結果が図 3 下段である。録音されている女性の音声がある程度大きいこともあり、発話区間全体で概ね F0 が推定できている。しかし細部を見ると、発話終盤の「だけど」の区間や発話最後の「が」の部分の F0 が得られていない。

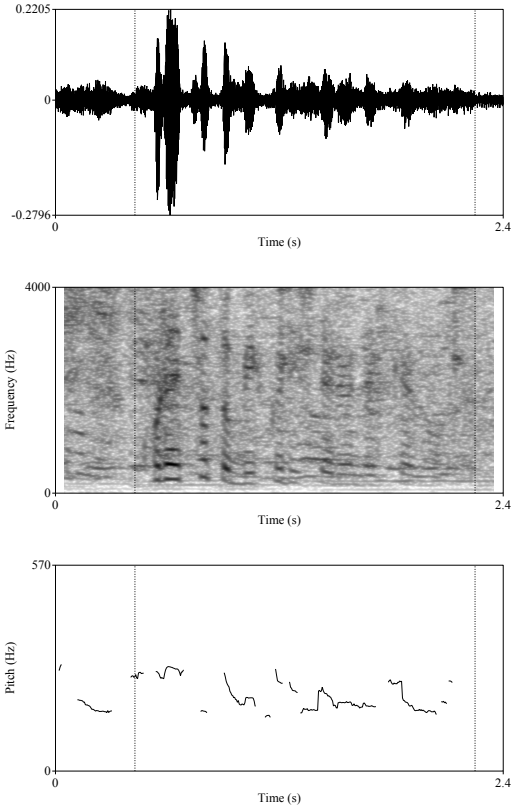



図 5　飲食店内の会話（会話 ID K002_018）における発話音声「これちょっとびっくりしてるんですけど」：（上）音声波形、（中）スペクトログラム、（下）推定 F0

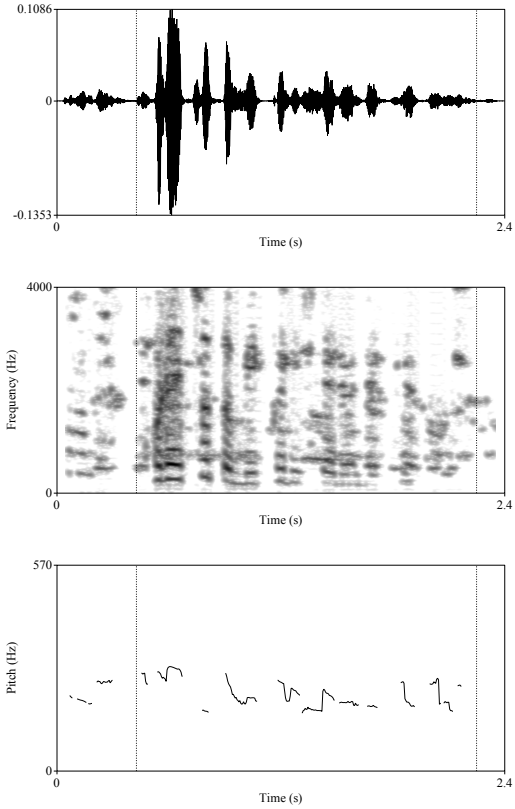



図 6　飲食店内の会話（会話 ID K002_018）における雑音抑圧処理後の発話音声

図 3 の音声信号に雑音抑圧処理を施した結果を図 4 に示す。図 4 上段の音声波形から明らかなように、やはり音声がない箇所では雑音がほとんど除去されている。図 4 中段のスペクトログラムでも無音区間では雑音がほとんど除去され、音声がある箇所でもスペクトル構造がより良く観察できるようになっていることがわかる。このような雑音抑圧の結果として、図 4 下段の推定 F0 が得られた。図 3 下段と比較すると、発話終盤の「だけど」や「が」の部分の F0 が得られている。特に最後の「が」の F0 は発話末の音調に関わる特徴であり、この部分の F0 が観察できることはコーパス構築の観点からも重要である。

次に、会話 ID K002_018 の飲食店内での女性の会話音声の一部を図 5 に示す。この飲食店では BGM として店内に音楽が流されている。そのため、図 5 上段の音声波形の点線区間内の「これちょっとびっくりしてるんですけど」という発声に、店内 BGM の女性ボーカルの音楽が入り込んで会話音声と重なっている。また、周囲の客のざわめき声も入っているため、発話区間外にも何らかの音が録音された状態になっている。図 5 中段のスペクトログラムからは、まず、周囲の客のざわめき声により、これまでの自動車走行音や機械音と同様に時間-周波数領域全体にエネルギーが存在していることがわかる。さらに発話区間の後半に重なるように、

ほとんど周波数が変わらず長く時間的方向に伸びたエネルギーが存在しているが、これは女性ボーカルの長く伸びた歌声によるものである。この歌声は時間とともに高さや大きさが変化することからざわめき声のような定常雑音とは異なる特性を持っている。図 5 下段を見ると、発話区間の前半部分では正しい F0 が取れているように見えるものの、後半部分では重畳された歌声の影響で発話がない箇所でも F0 が推定されていたり同じ高さの F0 値が連続している箇所が見受けられ、正確な F0 推定が行われているとはいえない。

図 6 はこの飲食店内の音声に雑音抑圧処理を施した結果を示している。図 6 上段や中段の音声波形やスペクトログラムからはこれまでの例と同じように雑音が除去されているように見える。実際に雑音抑圧処理後の音声を聴取すると、ミュージカルノイズが大きいものの確かにざわめき声は抑制されている。しかしながら、BGM の歌声はまだ聴き取れる状態になっている。そのため、図 6 下段の推定 F0 で発話中の無音区間になるべき箇所に着目すると、発話区間中盤の F0 値が水平に現れている箇所のように歌声の影響による F0 が推定されてしまっており、雑音抑圧処理では完全には対応できていないことがわかる。

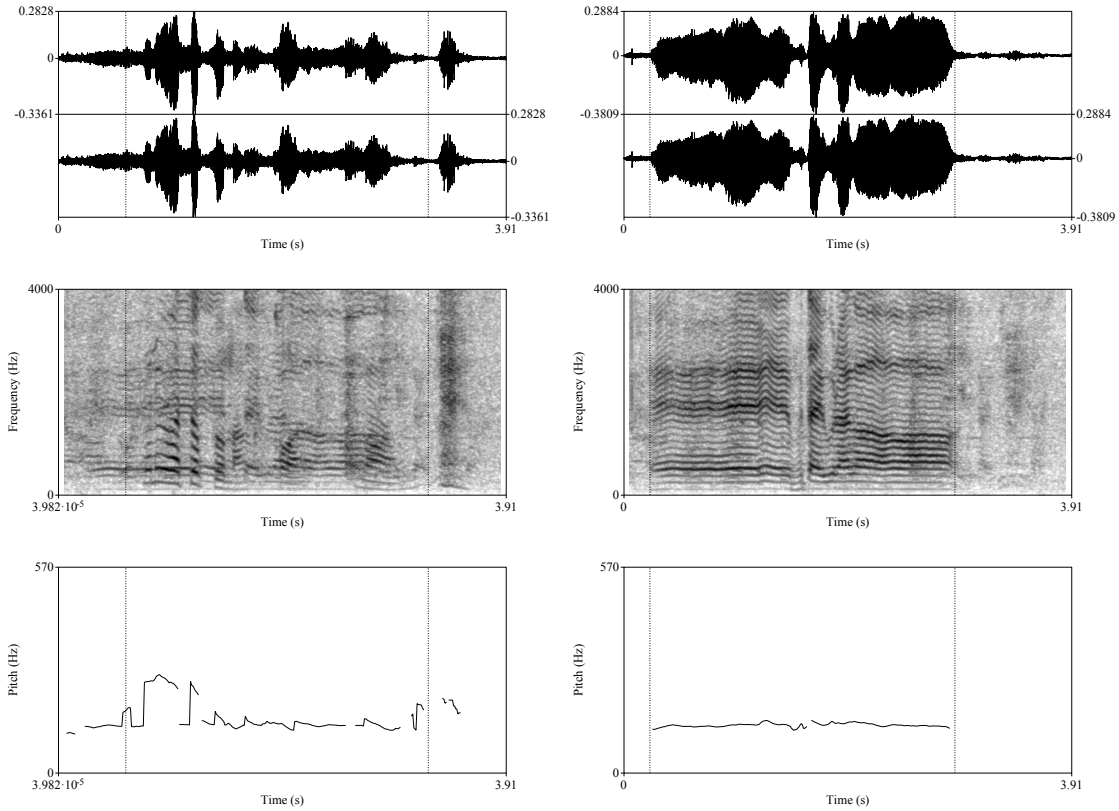
以上をまとめると、録音された日常会話音声に雑音が含まれていても、その雑音が周波数の時間的な変化が小さい、すなわち高い定常性を持つならば、雑音抑圧処理を行うことで F0 推定の精度を向上を期待できる。しかし、定常性の低い雑音に対しては、今回用いたような雑音抑圧処理ではその影響を完全に排除することができず、特に図 5 に示したように人の声同士が重なった状態の雑音抑圧は困難である。

### 2.2 発話の重複

CEJC では日常生活の活動を妨げない範囲で可能な限り話し相手の声が混入しないように収録方法を考慮しているが、発話者の口元とマイクがある程度離れているため、同時に発声された音声のそれぞれが十分に聴き取れる状態で録音されている場合がある。

図 7 に、飲食店での会話（会話 ID T001_014）の中から発話が重複して録音されている例を示す。この会話データは、飲食店でよくある周囲の客のざわめき音が入っているものの、それぞれの話者の音声が相対的に大きく録音されているため比較的 F0 が推定しやすい状態にある。しかし、話者同士の距離が近いこともあり、相手の音声が入りやすい状況になっている。

図 7(a) 上段は点線区間内で話者『師匠』が「それはちょっとなんかこう まあ」と発声している録音音声の波形を示している。この場面では師匠の発話とほぼ同じタイミングで師匠の正面に座っている話者『奥村』が「へー舞台裏側は:」と平坦な抑揚で発声している。その録音波形を図 7(b) 上段に示す。奥村が装着していた IC レコーダには奥村の発話が十分に大きく録音されており、周囲のざわめき音や同席している師匠の音声はほとんど入り込んでいない。そのため、図 7(b) 中段のスペクトログラムでは雑音の影響がほとんどない状態で奥村の発話の綺麗な調波構造が観察できる。また、図 7(b) 下段に示すように、F0 推定も問題なくできる。一方、師匠が装着している IC レコーダにはこの奥村の発話音声も入り込んでおり、師匠の発話と奥村の発話が重複した状態で録音されている。図 7(a) 中段のスペクトログラムを見ると、師匠の発話の周波数成分に加えて図 7(b) 中段に見られる奥村の発話の周波数成分も表示されていることがわかる。この録音音声から F0 推定を行った結果が図 7(a) 下段であるが、奥村の



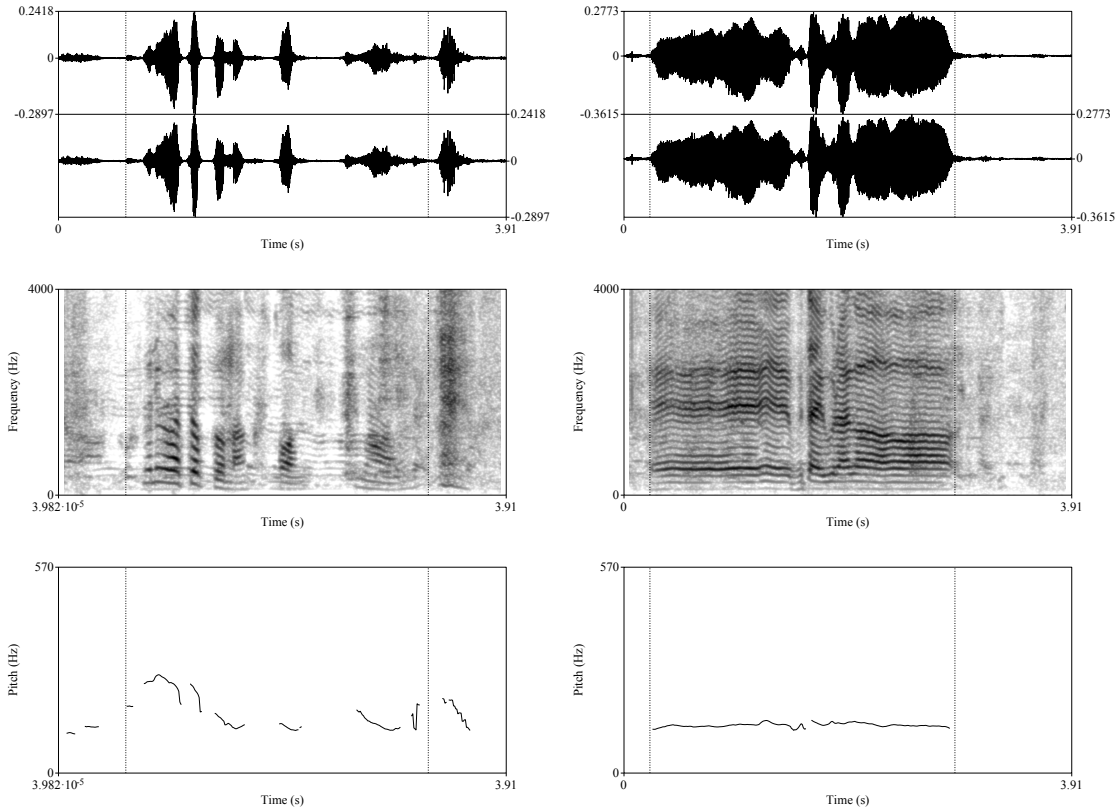

(a) 師匠「それはちょっとなんかこう まあ」　　(b) 奥村「へー舞台裏側は:」

図 7　飲食店での会話（会話 ID T001_014）において重複して録音された発話音声:（上）音声波形、（中）スペクトログラム、（下）推定 F0

発話の F0 が大部分を占めており、師匠の発話の F0 はほとんど得られていない。このように、発話が重複されて録音された場合は、目的の発話の F0 を得ることが困難になる。また、発話音声は非定常な特性を保つため、前節の雑音抑圧処理により重複の影響を軽減することができない。

そこで、音源分離処理により、重複して録音された音声からそれぞれの発話音声を取り出すことを考える。本稿では、音源分離処理も条件や目的に応じて様々な手法が提案されているが、信号のスパース性を利用し複数チャンネルで観測された信号から音源のモデルパラメータを推定することで時間－周波数領域の情報をそれぞれの音源へと振り分け分離する一手法 (Ozerov and Bimbot 2012) を用いることとする。この手法は “Flexible Audio Source Separation Toolbox” (FASST) として C++ で実装され、Python や MATLAB から実行できるように無料で公開されており（関連 URL 参照)、容易に効果を確認することが可能である。

図 7 に示した音声に音源分離処理を施した結果を図 8 に示す。奥村の発話音声（図 8(b)）に関しては、音源分離処理後も音声波形、スペクトログラム、推定 F0 にほとんど変化は生じていない。一方、奥村の音声が重複していた師匠の発話音声（図 8(a)）を見ると、音源分離処理前と大きく異なっていることがわかる。図 8(a) 上段の音声波形では奥村の発話の波形が取り

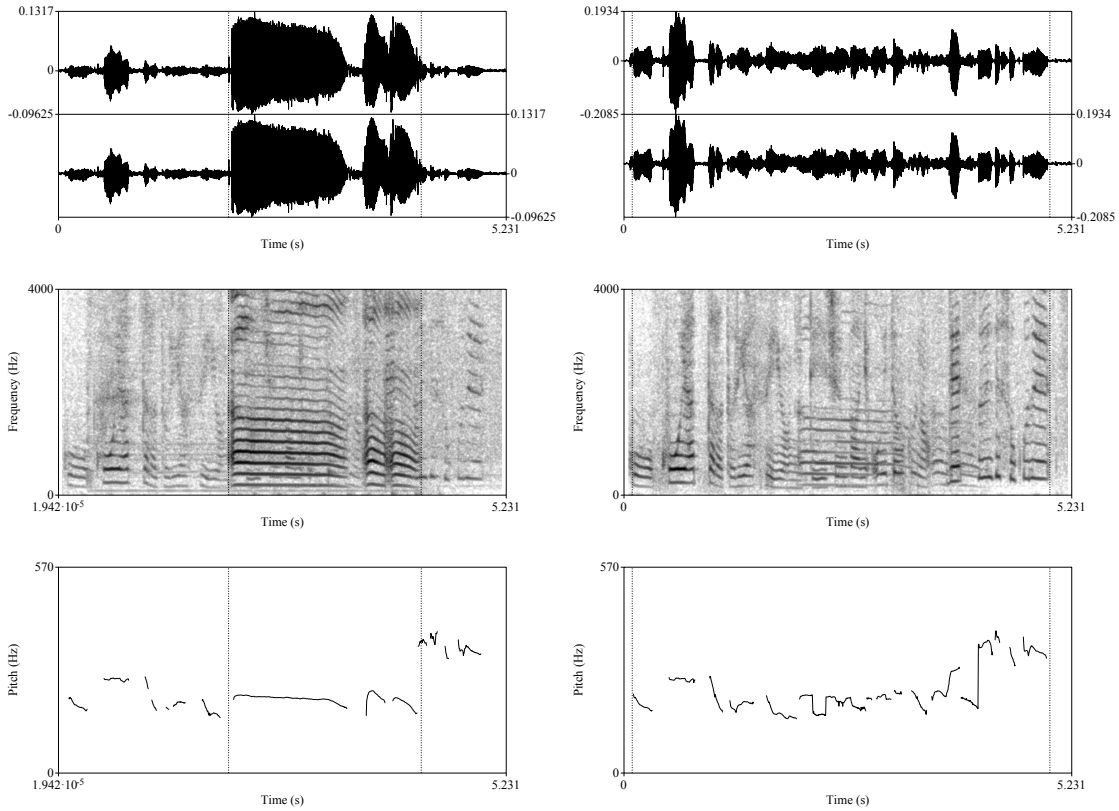



(a) 師匠「それはちょっとなんかこう まあ」　　(b) 奥村「へー舞台裏側は:」

図 8　飲食店での会話（会話 ID T001_014）における音源分離処理後の発話音声:（上）音声波形、（中）スペクトログラム、（下）推定 F0

除かれたことにより、師匠の発話区間が明確に判別できるようになっている。図 8(a) 中段のスペクトログラムでも奥村の発話音声のエネルギーが大幅に軽減されていることが見て取れる。この音声波形から推定した F0 は図 8(a) 下段に示すように、奥村の発話の F0 は全く見られず、師匠の F0 が観察できるようになった。このように重複した音声であっても、目的音声の F0 推定のために音源分離処理が活用できる。

ただし、常に音源分離処理だけで解決するわけではない。図 9 に、住宅内での会話（会話 ID C001_002）の中から発話が重複して録音されている箇所を一つ示す。この場面では女性 2 人が向かい合って座っており、話者『美沙』が「こうやったら取りやすいようにってゆう順番に置いておくの で それでそこで」と話している途中で話し相手の『玲子』が「うーん うんうん」と相槌を打っている。住宅内の静かな環境で会話していることもあり、周囲の雑音は全く入っていないが、お互いが装着している IC レコーダの録音に相手の発話音声が入り込んでいる。そのため、図 9(a) 上段の音声波形では、点線区間内の「うーん うんうん」と言う発話の前後に美沙の発話も入っていることがわかる。他方、図 9(b) の音声波形ではわかりにくいものの、中段のスペクトログラムを見ると美沙の発話の中央付近に玲子の相槌のエネルギーが表示されており、やはり玲子の発話が入っていることがわかる。ここで図 9(a) 下段の推定 F0 見ると、



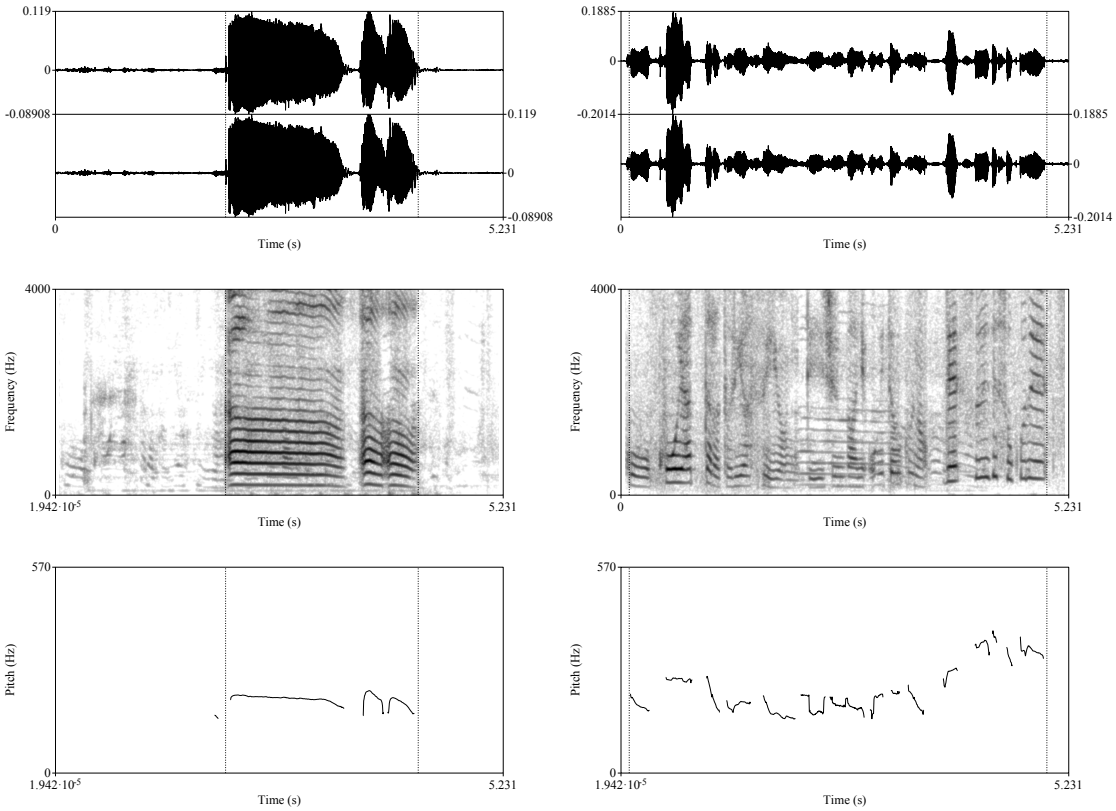

(a) 玲子「うーん うんうん」

(b) 美沙「こうやったら取りやすいようにってゆう順番に置いておくの で それでそこで」

図 9　住宅内での会話（会話 ID C001_002）において重複して録音された発話音声:（上）音声波形、（中）スペクトログラム、（下）推定 F0

玲子の発話の前後に美沙の発話の F0 が現れている。また、図 9(b) 下段からは美沙の発話の中央付近で玲子の相槌に対応する F0 が現れており、お互いに発話音声が重複して録音されていることから、推定 F0 がそれぞれ相手の音声の影響を受けてしまっている。

図 9 に示した音声に音源分離処理を施した結果を図 10 に示す。図 10(a) では、玲子の発話音声だけを取り出すことができており、対象としている発話の F0 だけが推定されている。しかし、図 10(b) では、美沙の発話の中にまだ玲子の発話が残っており、推定 F0 にも玲子の発話音声の影響が以前として現れている。この原因として、玲子の発話を音源としたモデルの推定がうまく働かなかったことが考えられる。今回使用した音源分離手法ではモデル推定にはその音源の音だけが存在する区間が必要であるが、会話 ID C001_002 では全体的に美沙の語りがほとんどであり、玲子だけが発話している箇所が少なかったため、モデル推定に用いる区間を適切に設定することができなかった可能性がある。この例のように、音源分離処理を効果的に働かせるためには条件があり、より正確に F0 を推定するためにはさらなる工夫が必要となる。




(a) 玲子「うーん うんうん」

(b) 美沙「こうやったら取りやすいようにってゆう順番に置いておくの で それでそこで」

図 10　住宅内での会話（会話 ID C001_002）における音源分離処理後の発話音声:（上）音声波形、（中）スペクトログラム、（下）推定 F0

## 3. おわりに

日常場面の会話が収録されている CEJC モニター版を基に、日常会話音声の録音データから F0 を推定する際に生じる周囲の雑音や発話の重複の影響を示した。また、そのような F0 推定に問題の生じる発話音声に対し、音声情報処理技術により前処理として雑音抑圧や音源分離を施すことで、正しく推定できなかった音声であっても F0 を得られる場合があることを示した。ただし、本稿で用いた手法により改善されない場合もあるため、音声の状態に応じた雑音抑圧・音源分離手法を検討する必要がある。

## 謝　辞

本研究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」および「大規模日常会話コーパスに基づく話し言葉の多角的研究」の成果である。

## 文　献

Hanae Koiso, Yasuharu Den, Yuriko Iseki, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, and Yasuyuki Usuda (2018). "Construction of the Corpus of Everyday Japanese Conversation: An Interim Report." *Proceedings of LREC2018*, pp. 4259–4264.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara (2000). "Spontaneous Speech Corpus of Japanese." *Proceedings of LREC2000* Vol. 2., pp. 947–952.

Paul Boersma, and Vincent van Heuven (2001). "Praat, a system for doing phonetics by computer." *Glot International*, 5:9/10, pp. 341–347.

Cyril Plapous, Claude Marro, and Pascal Scalart (2006). "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement." *IEEE Transactions on Audio, Speech and Language Processing*, 14:6, pp. 2098–2108.

Alexey Ozerov, and Emmanuel Vincentand Fr'ed'eric Bimbot (2012). "A General Flexible Framework for the Handling of Prior Information in Audio Source Separation." *IEEE Transactions on Audio, Speech, and Language Processing*, 20:4, pp. 1118–1133.

### 関連 URL

Praat: doing phonetics by computer　`http://www.fon.hum.uva.nl/praat/`

『日本語日常会話コーパス』モニター版

`https://www2.ninjal.ac.jp/conversation/cejc-monitor.html`

Wiener filter for Noise Reduction and speech enhancement - MATLAB Central

`https://jp.mathworks.com/matlabcentral/fileexchange/24462-wiener-filter-for-noise-reduction-and-speech-enhancement`

Flexible Audio Source Separation Toolbox　`http://fasst.gforge.inria.fr`