Vol.2011-CH-92 No.4
2011/10/8

『日本語話し言葉コーパス』を用いた自発音声の分析

前川喜久雄†

従来の音声研究では、実験的に統制されていない自発音声の研究はほぼ不可能と考えられてきた。しかしアノテーションを施した大規模なコーパスがあれば自発音声も分析可能になる。自発音声の分析結果によって従来の定説が覆されたり、実験環境下で収録された朗読音声の分析では発見が困難な韻律現象が発見されたりすることを、X-JToBI でラベリングされた『日本語話し言葉コーパス』コアの分析事例に基づいて指摘する。

Analysis of Spontaneous Speech using the *Corpus of Spontaneous Japanese*

Kikuo Maekawa†

Experimental study of spontaneous speech (i.e. the real speech that is not controlled by experimenters) has long been regarded to be impossible by many researchers. It turned out recently, however, that the study of spontaneous speech can be successful given a large-scale annotated corpus of spontaneous speech. In this paper, I will present some results of the analyses of the X-JToBI annotated part of the Corpus of Spontaneous Speech known as the CSJ-Core. These results include the cases where traditional analyses of Japanese phonetics were shattered by the analyses of spontaneous speech. There are also cases where analyses of spontaneous speech lead to better understanding of prosodic phenomena that are hardly observable in an experimental setting.

1. はじめに

自発音声(spontaneous speech)は実験者による実験的な統制を受けていない自然な発話を意味しており、朗読音声(read speech)ないし実験音声(laboratory speech)の反対概念である。音声研究の本来的な対象は自発音声であると考えられるのだが、実際には過去の音声研究で自発音声が分析対象とされることは稀であり、実験的な統制を受けた朗読音声の分析が音声研究の主流（というよりもほぼすべて）を占めてきている。

朗読音声が音声研究の主要な対象とされたのは、音声研究が自然科学の一部として自然科学の研究方法に従ってきたことの結果である。実験において刺激を厳密に統制することは、科学的方法のイロハとして受容されてきたが、自発音声には、その定義上、実験的な統制を施すことができないのである。

しかし音声が情報伝達の手段であることを考えるとき、朗読音声にのみ依拠した実験的音声研究で音声の本質が十分に解明されるとは考えにくい。人間の言語コミュニケーションにおいては､分節音素の語彙的対立に代表される主知的な言語情報以外に、主に韻律特徴によって伝達されるパラ言語情報や非言語情報が豊富に伝達されていることが知られている（ちなみにこれらの情報の大部分は書き言葉からは脱落する）[1]。

ここで重要なことは、これらの情報の表出は、必ずしも話者によって意識されていないという点である。例えば日本語（東京語）の句末・文末に生じる上昇イントネーションには少なくとも4種類の音声的な変種が存在していることが知られているが[2]、これをランダムに選ばれた（それが統計処理の前提である）被験者に、予め決められたフレーズにのせて発声させることは著しく困難である。発話に伴う意図（例えば反問）を説明する等の方法によって擬似的な発話データを収録することは不可能でないが、そのようにして得られたデータがその被験者が特定の社会的条件下で実際に生成する上昇イントネーションの変種の特徴を正確に反映したものとなっているかどうかは保障の限りでない。

このような問題が典型的に生じるのは、イントネーションに代表される、いわゆる句レベルの韻律特徴（アクセントは語レベルの韻律特徴）の研究においてであるが、本稿でも示すように、子音や母音などの分節音素の変異の研究でも同様の問題が生じることが知られている。

このような、従来の実験的研究では十分に把握することが困難な音声現象を科学的に研究するためには､実際に用いられた自発音声をそのまま分析対象とするしかなく、実験的統制を受けていない自発音声の分析から或る程度の信頼性をもった結論を引き出すためには、大量のデータを分析するしか方法がない。以下本稿では『日本語話し言葉コーパス』を利用して筆者が実施した自発音声研究の成果を報告する。

† 国立国語研究所言語資源研究系
Dept. Corpus Studies, National institute for Japanese Language and Linguistics

2. データ

2.1 『日本語話し言葉コーパス』

本稿では『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese)を分析する。『日本語話し言葉コーパス』は日本語の自発音声に関する最大のコーパスであり[3]、2004 年の公開以来、音声情報処理の領域を中心に広い領域で利用されている。現在までに CSJ を利用した学術論文が 500 件以上、博士論文が 10 篇以上執筆されており、日本語の自発音声コーパスとして代表的な存在と言ってよい。

CSJ の特徴はその規模（752 万語、660 時間）とともにアノテーションの豊富さにあるが、特に CSJ-Core と呼ばれるサブセット（50 万語、44 時間）には、X-JToBI と呼ばれるアノテーション方式に従った精密な分節音・韻律ラベリングが施されている。以下本稿で分析するのはこの CSJ-Core である。表 1 に CSJ-Core の簡単な仕様を示す。

表 1　CSJ-Core の仕様

講演種別	ファイル数	話者数(男/女)	総時間数
学会講演	70	24/46	18.7h
模擬講演	107	54/53	19.9h
対話	18	9/9	3.7h
再朗読	6	3/3	2.1h

講演種別のうち学会講演は理工学、人文科学、社会科学にまたがる各種学会での研究発表のライブ録音であり、模擬講演は人材派遣会社から派遣された年齢と性別を可能な限りバランスさせた話者による一般的な話題（私の住んでいる町、人生で最も嬉しかったこと、最近の出来事についての意見、等々）についてのスピーチである。これらのモノローグが CSJ-Core の大半をなすが、対比のために対話音声（学会講演ないし模擬講演の内容に関するインタビュー）と再朗読音声（学会講演ないし模擬講演を転記したテキストの同一話者による朗読）も数時間収録されている。CSJ-Core の話者は東京ないしその近郊出身でいわゆる標準語の話者である。

2.2 X-JToBI

CSJ-Core の全音声には CSJ 全体に提供されている形態論情報（短単位、長単位の二重解析）や節境界情報などに加えて、X-JToBI によるアノテーションが施されている。これは朗読音声用に開発された J_ToBI システムを自発音声用に拡張したものである[4]。X-JToBI ラベルは「単語層」「分節音層」「トーン層」「BI 層」「プロミネンス層」「注釈層」から構成されている。単語層ラベルは発話の構成する語（短単位）境界と語を構成する音素情報を提供している。分節音層ラベルは発話を構成する分節音（子音、母音、ポーズ等）と持続時間の情報を提供している。トーン層ラベルは A&M 理論[5]に基づいてイントネーションを音韻的なトーンの連鎖として表現している。BI 層ラベルは発話の韻律構造境界の相対的強度を表現している。そしてプロミネンス層ラベルと注釈層ラベルはトーン層ないし BI 層ラベルとの組み合わせによって、韻律現象の様々な変異に関する情報を提供するとともに、アノテーション上の問題点に関する情報も提示している。

3. 分節音の分析

3.1 ザ行子音の調音様式

CSJ-Core の分析例として最初にザ行子音/z/の調音様式の変異に関する分析結果を示す[6][7]。現代日本語の/z/は歯茎有声摩擦音[z]、歯茎有声破擦音[dz]のいずれでも発音されるが、先行研究の多くはこれを語頭位置では破擦音[dz]、語中（語頭以外）では摩擦音[z]という条件変異とみなしている [8][9]。この分析の妥当性を CSJ-Core の学会講演（56 講演）と模擬講演（106 講演）に生じた 14603 個の/z/の分析によって検討した。

最初に形態論的ないし韻律的単位の冒頭における破擦音の生起率（破擦率）を調査した。データ全体での破擦率は 35%であり、これが比較のベースラインである。

破擦率は短単位頭で 51%、長単位頭で 58.3%、アクセント句頭で 63.7%であった。いずれの単位においても語頭位置では語中位置よりも顕著に破擦率が上昇しているが、最高値をとるアクセント句頭でも 7 割に達しておらず、通常の意味での条件異音とはみなし難い。一方/z/の形態論的、韻律的な位置によらず、ポーズの直後に位置する場合を検討すると破擦率が顕著に上昇し 80%に達していた。また直前の分節音の影響を検討すると、/z/が促音ないし撥音の直後に位置する場合に破擦率が顕著に上昇していた（促音 73.7%、撥音 60.2%）。

これらの事実は、破擦率が/z/の調音運動に利用可能な時間の絶対値によって影響されているという仮説を示唆していると思われる。直前がポーズであれば、/z/の調音に時間的な余裕が生じるのは当然であるが、直前が促音や撥音の場合も、これらのモーラ音素には調音位置の情報が指定されておらず後続する子音（すなわち/z/）と一体化した長子音として調音されるために、/z/の調音には通常よりも長い時間をかけることが可能になる。

この仮説の妥当性を検討するために TACA(Time Allotted for Consonant Articulation)という量を定義する。TACA は/z/の持続時間を基本とし、/z/の直前に促音ないし撥音が位置していればその持続時間を追加した量である。/z/の直前にポーズが生じている場合にも/z/の持続時間の 2 倍を上限としてポーズの持続時間を加えることとし、促音ないし撥音とポーズがともに生じていれば両者の持続時間をともに追加する。このよ

うに定義された TACA と/z/の破擦率の関係は単調増加の関係を示し、TACA が 20ms から 240ms まで変動するにつれて、破擦率は 5%から 95%まで上昇し続けることが判明した（後掲する図 1 参照）。ロジスティック単回帰分析を行うと、TACA の値を知ることによって/z/の調音様式は 74%の精度で予測できる。

この分析結果は語頭位置にあっても TACA が小さければ破擦音は生じにくく、反対に語中位置にあっても発話速度が低下したり、モーラ音素の直後に位置することによって TACA が大きい値をとれば破擦音が生じやすいことを示しており、従来定説とされてきた条件異音説を否定するものである。このような結論を得ることができたのは、CSJ の音声には大幅な発話速度の変動が伴っており、そのため TACA の値も大幅に変動していたことによる。朗読音声でこの条件を再現することは非常に困難であろう。

3.2 /b, d, g/の閉鎖調音の弱化

/z/における破擦音と摩擦音のゆれと類似した変異は日本語の有声破裂音/b/, /d/, /g/にも生じている。これらの音素の調音ではしばしば声道の閉鎖が弱化して有声摩擦音として発音されることがある（IPA の記号を用いれば[β], [ð], [ɣ]である）。これら有声破裂音の閉鎖調音の弱化現象もまた TACA によって説明することができる[10]。

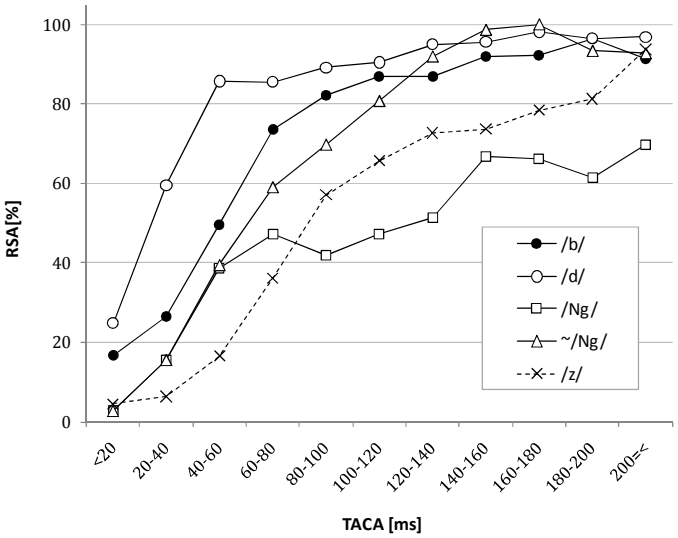

図 1 TACA と/b/, /d/, /g/, /z/における声道閉鎖率

図 1 は横軸に TACA の値（単位はミリセカンド）をとり、縦軸には/b/, /d/, /g/が声道閉鎖を伴う破裂音として実現される率（破裂率）ないし/z/がやはり声道閉鎖を伴う破擦音として実現される率（破擦率、単位はともに%）を配して、両者の関係を示したものである。/b/, /d/, /g/のいずれにおいても TACA と破裂率の間にはほぼ単調な増加関係が成立していることがみてとれる。

ただし、ここで/g/のデータは/Ng/と~/Ng/の２クラスに分けて表示されている。これは日本語では撥音/N/の直後の/g/はいわゆる鼻濁音（IPA の[ŋ]）として発音されることが多いからである。撥音直後の/g/（図の/Ng/）の破裂率は TACA が増大しても約 70%までしか上昇しない。これに対して撥音の直後にない/g/（図中の~/Ng/）の破裂率は TACA の増大につれてほぼ 100%近くまで上昇し続けることがわかる。

この図からはもうひとつ興味深い事実が読み取れる。それは TACA の値が比較的に小さい図の左半分においては、有声破裂音が所与の破裂率を達成するのに必要な TACA の値が音素によって異なっており/d/</b/</g/の関係をなしている点である。この関係には言語学上の意味があると思われる。図 2 に示されているように、/d/の調音位置である歯茎(alveolar)ないし歯(dental)においては、/d/に加えて有声摩擦音/z/と鼻音/n/の合計 3 音素が対立をなしているのに対して、/b/の調音位置である両唇(bilabial)では/b/以外には鼻音/m/が対立をなすのみであり、/g/の調音位置である軟口蓋(velar)においては/g/が唯一の音素である（鼻濁音[ŋ]は/g/の異音であり対立をなさない）。

図 1 における/d/</b/</g/の関係は、その調音点において対立する音素数を反映するもので、多くの音素が対立する調音点では音素の対立を確保するために少ない TACA の値でも閉鎖調音が実行され、反対に音素の対立が少ない環境では閉鎖調音も相対的に緩やかに実行されるのだと考えられる。調音運動の精確さの規準のひとつが言語的に要求される音韻対立の複雑さにあることは多くの音声学者によって夙に示唆されてきているが（例えば[11]）、それを実証したデータは少ない。図 1 はその珍しい例と言えるものであろう。

4. イントネーションの分析

4.1 PNLP

CSJ-Core を利用した韻律現象の分析をふたつ紹介する。ひとつは PNLP(Penultimate Non-Lexical Prominence)と呼ばれる現象の生起要因の分析である[12]。日本語の句末イントネーションには上昇下降調と呼ばれるものがある。これは発話の最終モーラ内部でピッチが上昇してピークに達したのち下降するもので、CSJ においても模擬講演を中心に頻繁に生じている。この上昇下降調の変種として、上昇のピークが発話の末尾から２モーラ目にずれているものがあり、これが PNLP である。PNLP がどのような条件によって生起するかは長年の謎であったが、CSJ-Core の分析によっていくつかの

重要な事実が判明した。

まず PNLP の生じている発話を対象に、PNLP の生起位置と生起確率の関係を検討すると、PNLP の生起確率は発話冒頭から次第に上昇し発話末から 2 番目のアクセント句において最高（例えば 5 アクセント句からなる発話の場合 50%程度）に達する。一方発話末のアクセント句に PNLP が生じることはほとんどない。

次にアクセント句数を単位として測定した発話の長さと各種の句末イントネーションの平均生起数との関係を検討すると、上昇イントネーションや（PNLP を除外した）上昇下降イントネーションは、発話長の増大につれて平均生起数も単調に増加するのに対して、PNLP の平均生起率は発話長が増大してもほとんど変動せずに 1.1 前後で一定している。

これらの事実から推測されるのは、PNLP は 1 発話には高々1 回しか生じず、PNLP が生じることによって発話の終了が予告されているという仮説である。この仮説を厳密に証明するのは今後の課題であるが、予備的な分析結果をみると、PNLP が生じた発話の直後では談話の話題が変化していることが多いように思われる[12]。

4.2 韻律特徴によるレジスターの判別

発話の韻律特徴は発話の種別によって大きく変動する。そのため例えばローパスフィルターをかけて分節的特徴を知覚不能にした音声を聴取してもその音声のレジスター（発話の目的などによって定まる発話種別）をある程度推測できると信じられている。この信念の当否を知るために、CSJ-Core に含まれる 201 ファイルのレジスターを韻律特徴のみによって判別することを試みた[13]。

X-JToBI で用いる 21 種のラベルのすべてについて 1 ファイル内における相対生起頻度情報を全ファイルを通して正規化した頻度情報および発話速度情報を従属変数として、表 1 に示されている 4 種のレジスター（発話種別）の線形判別を実施した結果、closed data で 85.1%、leave-one-out 交差評価で 78.1%の高い正判別率が得られる。また 21 種の X-JToBI ラベルのうち 15 種にはレジスターを要因とする一元配置分散分析で有意差が認められた。またステップワイズ法による変数選択を行うと 9 個のラベルが選択され、それらのラベルのみを用いて線形判別関数を構成すると、全ラベルを用いた場合に劣らない正判別率が得られることが判明した。分散分析で高い有意性を示したラベル、ステップワイズ法で選択されたラベルの大部分は、オリジナルの J_ToBI には含まれておらず、X-JToBI への拡張時に追加されたラベルが占めていたことから、X-JToBI の有用性が確認できた。

また各ファイルの冒頭から 60 秒ずつの音声を切りだしてその区間に含まれる X-JToBI ラベルのみを用いて線形判別分析を実施したところ、正判別率はファイルの冒頭の 1,2 分のデータを用いた場合にはやや低下し、反対にファイルの末尾 1,2 分のデータではやや上昇することが確認されたが、それ以外の位置のデータを用いた場合には、おしなべて closed data で 75%前後、交差評価で 70%前後の正判別率が得られることが分かった。60 秒という比較的少ないデータを用いた場合にも比較的よい結果が得られることがわかると同時に、レジスターの差異を示す韻律特徴はファイルの全体にわたって分布している特徴(versatile な特徴)であることがわかる。

5. おわりに

本稿では『日本語話し言葉コーパス』を用いて筆者自身が近年実施した自発音声研究の成果をかいつまんで紹介した。これ以外の成果については文献[14]を参照されたい。結論として CSJ-Core のように、ある程度大規模で、幅広いレジスターの音声を収録したアノテーション付コーパスがあれば、自発音声の研究は十分に可能であることが確認できた。また自発音声の分析によって、従来朗読音声の分析結果に基づいて提唱されてきた定説が覆されることがありうることが示された。今後は CSJ の分析を継続するとともに CSJ がカバーしていないレジスターの自発音声にも分析の手を広げたいと考えている。

参考文献

[1] 前川喜久雄・北川智利「音声はパラ言語情報をいかに伝えるか」認知科学, 9(6), pp.46-66, 2002.
[2] 川上蓁「文末などの上昇調について」国語研究, 16, pp.25‐46, 1963.
[3] 前川喜久雄「『日本語話し言葉コーパス』の概要」日本語科学, 15, pp.111-133, 2004.
[4] K. Maekawa, H. Kikuchi, Y. Igarashi and J. Venditti. “X-JToBI: An extended J_ToBI for spontaneous speech", *Proc. ICSLP2002*, Denver, pp.1545-1548, 2002.
[5] R. Ladd *Intonational Phonology*. Cambridge Univ. Press, 1996.
[6] K. Maekawa."Coarticulatory reinterpretation of allophonic variation: Corpus-based analysis of /z/ in spontaneous Japanese." *Journal of Phonetics*, 38(3), pp.360-374, 2010.
[7] 前川喜久雄「/z/の調音様式の変異」国語研プロジェクトレビュー, 5, pp.21-45, 2011.
[8] 服部四郎『音聲學』岩波書店, 1951.
[9] 天沼寧・大坪一夫・水谷修『日本語音声学』くろしお出版,1978.
[10] 前川喜久雄「日本語有声破裂音における閉鎖調音の弱化」音声研究,14(2), pp.1-15, 2010.
[11] Lindblom, Björn “Explaining phonetic variation: A sketch of H&H theory.” In W. J. Hardcastle and A.Marchal (eds.) *Speech production and speech modeling*. Dordrecht: Kluwer Academic Publishers, 1998.
[12] 前川喜久雄「PNLP の音声的形状と言語的機能」音声研究,15(1), pp.16-28, 2011.
[13] K. Maekawa. “Discrimination of Speech Registers by Prosody.” Proc. ICPhS 2011, Hong Kong, pp.1302-1305, 2011.
[14] 前川喜久雄『コーパスを利用した自発音声の研究』東京工業大学情報理工学研究科学位論文, 2011.