日本語自発音声韻律ラベリングスキーム *X-JToBI* の能力検証

Accuracy of Prosodic Labeling of Spontaneous Speech by *X-JToBI*

菊池英明 †‡ *　　前川喜久雄 †
KIKUCHI Hideaki　　MAEKAWA Kikuo
† 国立国語研究所　　‡ 早稲田大学
†The National Institute for Japanese Language　　‡Waseda University

Abstract: We are planning to give prosodic labels to about fifty hours of spontaneous Japanese speech. By our preliminary experiment, inter-labeler reliability of conventional labeling scheme, which was established for labeling read speech get worse in labeling spontaneous speech. To solve this problem, we have proposed the new prosodic labeling scheme named X-JToBI, the extended version of J_ToBI which has grown out of our work on annotating prosodic features of spontaneous speech. Describing prosodic feature more accurately is one purpose of proposition of the new prosodic labeling scheme, and improvement of inter-labeler reliability is another. In this paper, the results of analyses of inter-labeler reliability on prosodic labeling, especially the comparison with the results of J_ToBI labeling will be discussed mainly. At first, in the tone labels of boundary pitch movement, kappa of X-JToBI is higher than one of J_ToBI. Also, in the tone labels of accents and phrasal tones, kappa of X-JToBI is higher. Exact match between the time-stamp of tone labels and the timing of physical events seem to help improvement of inter-labeler agreement. Second, in labels of all break indices, kappa of X-JToBI is higher than one of J_ToBI. But there is no significant difference if excluding the labels of disfluency and fillers. This means that newly defined break index labels for disfluency and fillers improve inter-labeler agreement of prosodic labeling of spontaneous speech. Observed rate of agreement rose to 88% from 75% totally. From these results, it turned out that our extensions are effective in improvement of inter-labeler reliability.

1　はじめに

近年、音声認識や音声合成などの工学的研究や、言語学や音声学、談話分析などの言語学的研究での利用を目的として、多種多様な音声言語コーパスの構築が各方面で行なわれている [1]。こうしたコーパスでは、音声データに加えて、書き起こし情報、形態論情報、音素情報が音声データに加えて提供されるのが一般的であるが、最近では韻律情報を利用したいという声も大きく、韻律情報を利用しやすい形で記述・表現するラベリング手法の確立が求められている [2]。

現在、我々は日本語話し言葉コーパス (Corpus of Spontaneous Japanese: CSJ) の構築を進めている。このコーパスでは、全体約 700 万語規模のデータのうち約 50 万語分のデータに対して韻律情報を付与する予定である [3]。日本語話し言葉音声の韻律情報をラベリングするために、まず日本語音声の韻律情報のコーディング手法である J_ToBI モデル [4] について、話し言葉音声に適用した結果を分析し、自発性の高い音声のラベリングにおいて種々の問題が生じることを明らかにした [5]。そのうえで、その際の考察に基づいて J_ToBI を拡張し、新たな韻律ラベリングスキーム X-JToBI(eXtended J_ToBI) を提案した [6]。

一般にラベリングスキームの策定においては、いかに正しく情報を記録できるかが重要であり、正確性や安定性、再現可能性等の客観的評価が望まれる。筆者らはこれまでに 10 時間を超える量の話し言葉音声に対して X-JToBI によるラベリングを実施し、従来よりも高い記述力、安定性を実感している。そこで、本稿では X-JToBI のラベリングスキームとしての能力検証の一環として、始めにラベルの出現頻度を調べ、そのうえで正確性と再現可能性に注目してラベリング精度を調べた結果を報告する。

*連絡先: 〒 359-1192 埼玉県所沢市三ヶ島 2-579-15 早稲田大学人間科学部, E-mail:kikuchi@waseda.jp

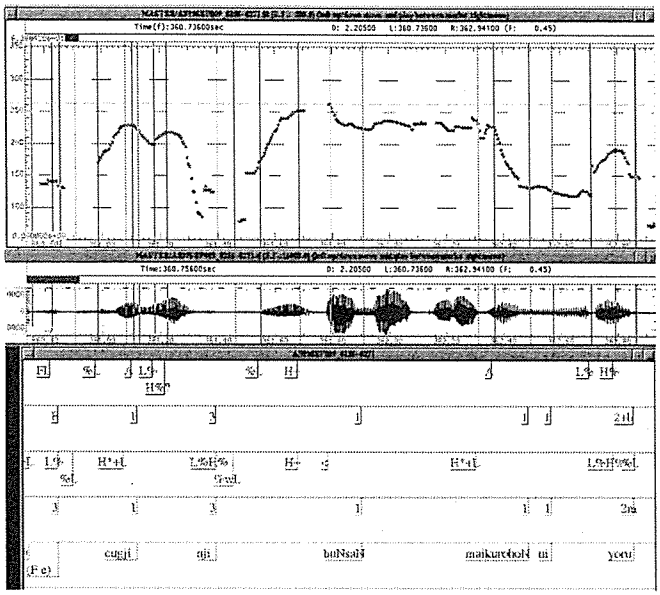

Fig. 1: 同一サンプルに対する J_ToBI および X-JToBI ラベリングの例
ラベル層は上から，X-JToBI の tone 層，BI 層，J_ToBI の tone 層，BI 層．
(「え次に分散マイクロホンによる」)

2 X-JToBI

まず始めに、日本語話し言葉音声の韻律ラベリングスキーム X-JToBI について、J_ToBI に対する拡張点を整理する。

(1) 分節音情報の組織的な付与を前提

従来の J_ToBI が tone ラベリングで要請していた各イベントの音韻論的位置の記述工程と物理的実現位置の記述工程のうち、前者はあらかじめ分節音情報を与えて tone ラベルの位置との物理的対応関係を明確にすることで省くことができる。

(2)Tone ラベルの拡張：物理的イベントの時間的対応

(1) の前提により、tone ラベルを物理的な実現位置に付与することで音韻論的位置の記述は不要になる。その結果、例えば「遅下がり」や「早上がり」を示す “>” や “<” のラベルが不要になる。原則として H は F0 の局所的極大値ないし変曲点、L は局所的極小値ないし変曲点をラベル位置とする。この原則にしたがえば、複合境界音調 (“L%H%”, “L%HL%”) は分解されたうえで H,L のそれぞれのイベントに対応付けられることになる。なお、自発音声において認められた新しい音調 (“L%LH%”) を句末境界音調カテゴリとして追加する。さらに、分節音持続時間の摂動増大によるピッチ保持に対して新たにエクステンダーと呼ぶ記号 (“>”) を導入する。

(3)Break Index ラベルの拡張：mismatch の分類

J_ToBI の運用では通常のイントネーションの構造と異なるために BI の判定に迷うケースが少なからず生じる。この場合に備えて J_ToBI には “p”,“-”,“m” の補助記号が準備されているが、これらを具体的にどのようなケースに適用するかの説明が不足しているために作業に支障をきたすことが多い。自発音声のラベリングではこの困難が極端に増幅される。X-JToBI では BI のインベントリを拡張するとともに、補助記号を整理して個々の用法を詳細化することに努める。

(4)Disfluency への対応

自発音声における種々の disfluency に対して、従来の J_ToBI で充分に表現できない語断片、語中のポーズ、フィラーをとりあげて、主に break index 層のインベントリを拡張することで対処する。

(5) その他音韻理論上の問題への対処

さらに、自発音声の韻律ラベリングには音韻理論上の典型的な問題がいくつかある。それらの問題に対処するために主に Miscellaneous 層 (一部 BI 層) のインベントリを拡張した。

同一の音声サンプルに対する J_ToBI および X-JToBI のラベリング結果を図 1 に示す。図では、冒頭のフィラー「え」に対して J_ToBI が一つのアクセント句としてとらえるのに対し、X-JToBI ではフィラー用のラベルで記述する。さらに、句末の複合境界音調に対して J_ToBI では複合的なラベル "L%H%" を付与するが、X-JToBI では物理的な変曲点と極大値にそれぞれ "L%" と "H%" を分けて付与する。

3 ラベル出現頻度

X-JToBI では自発音声に特有の現象を記述するためにいくつかの新たなラベルを導入している。ラベリング精度を測る前に、まず始めに実際の自発音声においてこれらがどの程度活用されるかを調べる。以下には、音韻的トーンを表す tone 層と単語間の韻律的区切りの度合を表す break index(以下、BI) 層にわけてラベルの出現頻度を分析した結果を示す。なお、分析の対象としたデータは、前述した CSJ の一部に対して一名の作業者が X-JToBI ラベリングを行なったものである。CSJ のデータは講演スタイルのモノローグが主であり、学会での研究発表の音声である学会講演と、スタジオで収録した一般人による模擬的な講演の 2 種類がある。ラベリングの対象としたデータにおける発話の総時間は模擬講演が 2.75(男性 0.99、女性 1.76) 時間、学会講演が 1.02 時間 (男性のみ) である。

3.1 Tone 層

先に 2 章 (2) で示した様に、X-JToBI では句末における局所的なピッチ変化 (boundary pitch movement, 以下 BPM) として従来の "L%H%", "L%HL%" の他に新たに "L%LH%" を導入した。表 1 に示した講演種別毎の BPM 出現頻度分布から、提案した "L%LH%" が少ないながらもどちらの種別にも出現していることがわかる。なお、講演種別の比較では模擬講演の方により多く現われているが、これは "L%HL%" の出現傾向からも模擬講演の方が学会講演より音調が多様であるためと考えられる。模擬講演における出現頻度分布の性差はほとんどない。

一方、tone イベントにおけるピッチ保持を表すために導入したエクステンダー (2 章 (2) 参照) は全ての tone ラベルのうち 1.85%の割合で出現していた。講演種別による差はほとんど見られなかったが、模擬講演において男性 1.62%、女性 2.40%と、やや女性の講演に多く出現していた。

3.2 BI 層

表 2 にはデータ種別毎の BI ラベル出現頻度分布を示す。なお、CSJ における韻律ラベリングで使用しないラベル ("1+w" など) は集計の対象から除外した。

この表において量の多少はあっても全てのラベルが利用されていることがわかる。特に、語断片に付与する "D" やフィラーに付与する "F" は、2 章 (4) で示した拡張に相当し、これらの出現頻度から (全体の 7.4%) 自発音声特有の現象を表現するのに役立っているといえる。フィラーのラベル ("F" および "<F") が学会講演でより頻出している以外は、講演種別による差はほとんど見られない。

なお、"PB" は「本当?と言いました」のように BPM を有する句が引用された場合などで文法的に許容されない BI が寄生的に発声していることを示すラベルである。このラベルもそれぞれ 0.06%、0.07%と低頻度ながら出現している。

表 1: BPM を表わす Tone ラベルの出現頻度分布
(() 内は各種別における BPM 総数に対する割合 [%])

tone ラベル	模擬講演		学会講演	
L%+H%	1683	(58.05)	1331	(78.80)
L%+HL%	1121	(38.67)	346	(20.49)
L%+LH%	95	(3.28)	12	(0.71)

表 2: BI ラベルの出現頻度分布
(() 内は各種別における BI ラベル総数に対する割合 [%])

BI ラベル	模擬講演		学会講演	
1	31696	(69.53)	8121	(55.73)
1+	42	(0.12)	9	(0.06)
1+p	254	(0.70)	164	(1.13)
2	4071	(7.82)	1333	(9.15)
2+	15	(0.03)	9	(0.06)
2+p	214	(0.41)	67	(0.46)
2+b	466	(0.90)	526	(3.61)
2+pb	48	(0.09)	8	(0.05)
3	7478	(14.37)	2617	(17.96)
D	280	(0.54)	97	(0.67)
D+	31	(0.07)	26	(0.18)
P	19	(0.04)	6	(0.04)
P+	12	(0.02)	5	(0.03)
<F	431	(0.83)	464	(3.18)
F	2426	(4.66)	1111	(7.62)
PB	61	(0.06)	10	(0.07)

4 ラベリング精度

本章では、ラベリング精度を調べた結果を報告する。ラベリング精度としては、暫定的に設定した正解との一致率によって正確性を、複数ラベラー間の一致率によって再現可能性を測定する。なお、ラベラー間一致率の指標としてはCohenが定義したκ[7]を用いる。κは、観測された一致率を$P(O)$、期待される一致率を$P(E)$とすると以下の式により得られる。

$$\kappa = \frac{P(O) - P(E)}{1 - P(E)} \quad (1)$$

ここでは、音声的に特徴のある平均約30秒程度の講演音声をCSJから9つ抽出して用意し、3名のラベラーがX-JToBIにしたがって韻律ラベリングを行なった結果を分析の対象とした。なお同じ音声データに対して多数のラベラーがJ_ToBIにしたがってラベリングを行なった[5]結果のうち暫定的に設定した正解に近い上位3名のラベルを比較対象として用いた。J_ToBIの3名のラベラーは過去にラベリングの経験があり、熟練度は比較的高いといえる。

ラベリングに先立って、語境界および語のアクセント情報を記すword層のラベルを形態論情報に基づいて作成し、BI層の初期値としてword層ラベルと同じ位置にBI=1を与えている。tone層の初期値は与えなかった。

4.1 BI層

まず、BI層について、正解ラベルに対する正解率をラベル種類毎に集計した結果を表3に示す。表中、J_ToBIとX-JToBIの対応するラベルを同一行に示している。この表から、サンプル数の少ない"1+"を除くほぼ全てのラベルにおいて正解率が向上していることがわかる。なお、アクセント句境界にBPMが生じることによって"2"よりも強い区切りになることを示す"2+b"の正解率が、主に対応するJ_ToBIの"2m"の正解率に比べて低くなっている。この原因として、「という」などの引用の形で「と」が高いピッチで発声されるような、アクセント句境界の位置判定が困難になるケースで、熟練したJ_ToBIの3名のラベラーの方がパターンとして処理できていることがあげられる。こうしたケースに対しては特に基準を明確にしたうえでラベラーの訓練を行なう必要がある。

次に、ラベラー間一致率を調べたところ、J_ToBIにおいて$\kappa = 0.64$ ($P(O) = 0.78$, $P(E) = 0.40$)であり、X-JToBIにおいては$\kappa = 0.73$ ($P(O) = 0.83$, $P(E) = 0.37$)であった。ラベルの種類を増やしたにもかかわらず一致率が向上していることから、本スキーム設計の有効性がうかがえる。

表3: BIラベルの正解率
(() 内は正解ラベルにおける総出現数)

J_ToBI			X-JToBI		
ラベル	正解率		ラベル	正解率	
1	91.3	(593)	1	94.9	(531)
2	74.0	(123)	2	74.4	(112)
3	70.5	(182)	3	75.1	(181)
2-	33.3	(1)	1+	0.0	(1)
1p	47.9	(16)	1+p	81.5	(9)
3-	–	(0)	2+	–	(0)
2m	80.5	(29)	2+b	66.7	(30)
2p	27.3	(11)	2+p	33.3	(8)
3m	0.0	(1)	2+pb	33.3	(4)
			D	83.3	(4)
			D+	44.4	(3)
—	—	—	<F	76.2	(7)
			F	91.0	(63)
			PB	33.3	(2)
全体	83.2	(956)	全体	86.1	(956)

なお、X-JToBIの正解として"D"、"F"が付与されている箇所を除いてラベラー間一致率を求めたところ、J_ToBI、X-JToBIにおいてそれぞれ$\kappa = 0.69$、$\kappa = 0.71$と差が小さくなることから、特に"D"と"F"の効果が大きいといえる。

4.2 Tone層

次に、tone層のうちBPMのラベリング精度を調べた。まず、正解に対してアクセント句境界が一致する箇所のみを対象に句末境界音調毎に正解率を調べた結果を表4に示す。

表4: BPMの正解率

(a)J_ToBI

	正解ラベル		
ラベリング結果	L%	L%H%	L%HL%
L%	144	15	11
L%+H%	7	57	6
L%+HL%	1	0	10
正解率 [%]	94.7	79.2	37.0

(b)X-JToBI

	正解ラベル			
ラベリング結果	L%	L%H%	L%HL%	L%LH%
L%	360	34	17	0
L%+H%	12	157	3	0
L%+HL%	5	7	40	0
L%+LH%	0	0	0	0
正解率 [%]	95.5	79.3	66.7	-