音声言語情報処理 7−10
（1995. 7. 21）

習熟による対話音声情報の書き起こし精度の定量的評価

小林 聡　　北澤 茂良
e-mail:{skoba, kitazawa}@cs.shizuoka.ac.jp
静岡大学 電子科学研究科
〒432 静岡県浜松市城北 3-5-1

あらまし 周辺言語的な特徴は多くのアスペクトを持っており、会話において重要な役割を担っている。会話の書き起こしは、その分析において有用な方法である。しかし、周辺言語的な特徴は書き起こし作業者の主観によって判断されるため、作業者間での書き起こし内容の不一致は避けられない。作業者間での不一致は、分析の正確さや客観性において問題となる。本実験では、作業者の技術やグラフィカル・インターフェースの使用、および書き起こしテキストを元に作成された合成音との比較による書き起こしテキストの検証といういくつかの条件において、作業者間のクロス・チェックにより書き起こしテキストの不一致について調べた。その結果、作業者の習熟およびグラフィカル・インターフェースの使用は、作業者間での書き起こしテキストの一致率の向上に効果が有った。

キーワード 非言語情報、書き起こし精度

Consistency of Inter-Transcribers' Transcription

Kobayashi Satoshi　　Kitazawa Shigeyoshi
Graduate School of Electronic Science and Technology,
Shizuoka University
3-5-1 Johoku, Hamamatu, Shizuoka, 432 Japan

Abstract Paralinguistic features have many aspects and play important roles in human dialogue. Transcription of dialogue is a useful technique for analysis of dialogues. However, paralinguistic features are classified by transcribers' subjective. Inconsistencies, such as omissions and misleading, between transcribers are unavoidable. Such inconsistencies between transcribers arise the problem about accuracy and objectivity of analysis. In this paper, we evaluated those errors and inconsistencies based on the cross-check between transcribers on conditions such as transcribers' skill, using graphical interface and verifying transcription with synthetic sounds. The results show training and graphic interface had positive effects on consistency of inter-transcribers' transcription.

Keywords nonverbal feature, consistency of transcription

1 はじめに

自然な会話において非言語コミュニケーションは重要な役割を担っている。その中には、発声法や声質などの周辺言語と呼ばれるものも含まれている [1]。周辺言語は多くのアスペクトを持っているが、それらのいくつかは感覚的なものであり測定が難しいものである。また他のものは音響的な特徴として測定可能なものである。これら声の大きさや発話速度などの周辺言語的特徴の書き起こしは会話の分析に有用である。しかし、それらは書き起こしにあたった作業者の主観によって分類および記述される。我々は、書き起こし作業者の技術の向上、グラフィカル・インターフェースの使用、および書き起こしテキストを元に作成された合成音との比較による書き起こしテキストの検証を行なった場合といういくつかの条件において、作業者間での書き起こしテキストの一致の程度を、クロス・チェックにより調べた。

2 実験条件

本実験には 2 つの音声データを使用した。1 つは被験者のトレーニング用であり、もう 1 つはテスト用である。音声データはいずれも女性の司会者と男性のゲストとの間で行なわれたインタビューであり、自然な会話である。長さは約 40 秒であり、44.1kHz、16bit、モノラル録音されている。聴取にあたっては、SparcStation 10 上で、ヘッドフォンを用いて行なった。

我々は、被験者を 2 つのグループに分割した。1 つは音声波形とピッチ波形を参照しながら書き起こし作業を行なう、3 名のグループである。このグループを W-group と呼ぶ。もう 1 つはグラフィカル・インターフェースを使用しないグループであり、4 名からなる。こちらを B-group と呼ぶ。クロス・チェックによって書き起こしテキストの比較を行なうため、それぞれのセッションにおいて W-group は 3 つ、B-group は 6 つの比較ペアがある。いずれの グループの被験者も、音声データの聴取にあたっては任意の部分を聴取可能である。

このような条件で、被験者は 5 回のセッションを行なった。第 1 から第 3 セッションでは、被験者はトレーニング用の音声データの書き起こしを行なった。第 4 セッションでは、テスト用データの書き起こしを行なった。第 5 セッションでは、被験者は自分の書き起こしテキストをもとにした合成音を聞き、オリジナルの音声データと比較することにより自分の書き起こしテキストを修正した。

書き起こしにあたっては、言語音および笑い声などはひらがなによって書き起こし、その他の非言語的特徴については TEI のタグと実体によって書き起こした [3]。被験者のうち、5 名は TEI のタグおよび実体を使用した書き起こしについて若干の経験が有り、残りの 2 名はまったくの未経験者である。被験者は全員、TEI の記述法および書き起こし作業の進めかたなどについて、作業開始前に講習を受けた。ここで、発話のとらえ方やその他の非言語的特徴の書き起こしの基準などを与えた。また、各セッションの終了時にも、被験者からの質問に答えるなどのためのミーティングを行なった。

本実験では、書き起こしテキストの不一致の評価を、被験者間のクロス・チェックによって行なった。これは、自然な会話においては台本などが無いこと、また不明瞭な発声が有ること、さらに本実験で使用した音声データはモノラル録音されているために発話のオーバーラップによって発話内容が不明瞭になるなどの理由により、正解とみなせるものが無いためである。さらには、本実験において問題としているのは、被験者間での書き起こしテキストの差異であるため、作業者間でのクロス・チェックにより一致率の計測を行なった。

今回、一致率の評価は、以下のような方法で行なった。

$$R = \frac{a+b}{A+B} * 100[\%]$$

A: クロス・チェックの片方において書き起こされた要素の数 (ひらがな、タグ、実体)

B: クロス・チェックのもう片方において書き起こされた要素の数

a: A の側でのクロス・チェックにおいてマッチした要素の数

b: B の側でのクロス・チェックにおいてマッチした要素の数

(a = b)

ここで、"A" はクロス・チェックの対象となっている書き起こしテキストのうちの片方に含まれている要素の数で有り、また"B"は、クロス・チェックの対象となっているもう 1 つのテキストに含まれている要素数である。ここで

の要素とは、ひらがなであったり、タグ、実体である。仮にタグの一致率の場合であれば、"A" や "B" はそれぞれの書き起こしテキストに含まれるタグの数になる。"a" や "b" も同様に数えられる。

3 結果

3.1 第1セッション

第1セッションでの被験者間における一致率を表 1に示す。表 1において、"合計"は、クロス・チェック中にカウントされた要素の総計である。"Char"はひらがなで表記された音について、また "Tag" と "Entity" はそれぞれ TEI タグ および実体で表記された非言語的特徴を現わす。"All" は、"Char"、"Tag"、"Entity" の合計である。タグは句などの比較的長い区間における特徴を表記するものであり、また実体は音節などの短い区間における特徴を表記するものである。ここで、同一もしくは類似する内容が、同一もしくは近隣に書き起こされていた場合に、それらの表記が一致すると考えての一致率である。

なお、ここで「類似する内容」とは、ひらがなでの表記が正書法に従ったものおよび音を記述してあるもの、書き起こされた音が他の音にも聞こえるなどと註釈が付けられているものや笑い声、また非言語的特徴においてその変化の程度が異なるだけのものである。例えば、助詞「は」が「わ」と書き起こされている場合と「は」と書き起こされている場合。書き起こしでは「あ」と書き起こされているものの、「「え」とも聞こえる」などと註釈が付けられており、クロス・チェックの相手方においてその音が「え」と書き起こされている場合。TEI タグ "tempo" の場合であれば、"aa" と "a" など。これらは類似した内容と考えて一致率の計算を行なった。

表 1において、"Char"は 90% の一致率を得ている。しかし、"Tag"と "Entity"においては、"Char"の 1/3 以下の一致率にとどまっている。また、W-group は B-group よりも高い一致率を得ている。

3.2 第4セッション

3回のトレーニング・セッションの後、このセッションでは被験者にテスト用の音声データを書き起こしてもらった。表 2に第4セッションにおける被験者間での書き起こしテキストの一致率を示す。

表 1: Consistency Rate in 1st Transcription[%].

	B-group		W-group	
	%	合計	%	合計
All	70.4%	7,938	75.5%	3,919
Char	88.4%	5,304	90.9%	2,702
Tag	36.0%	2,385	42.5%	1,169
Entity	14.5%	249	16.7%	48

表 2: Consistency Rate in 4th Transcription[%].

	B-group		W-group	
	%	合計	%	合計
All	75.5%	7,161	80.0%	3,770
Char	91.6%	4,854	92.1%	2,438
Tag	44.9%	2,094	58.8%	1,278
Entity	10.3%	213	31.5%	54

表 2において、"Char"は 91% 程度の一致率を得ている。しかし、"Tag"と "Entity"は、その 1/2 以下程度である。また、W-group は B-group よりも高い一致率を得ている。

この表 2の、B-group の "Entity"を除くすべてが表 1よりも高い一致率を得ている。

3.3 第5セッション

このセッションでは、被験者は自分の書き起こしテキストをもとにして非言語情報を再現した合成音を聞くことによって、自分の書き起こしテキストの修正を行なった。

書き起こしテキストを元に合成音を作るプログラムを作成した。このプログラムはいくつかの TEI タグと実体を解釈可能であり、その内容を表 3に挙げる。このプログラムは正弦波音によって、音節に対応する単位の音を合成する。それぞれの単位は、書き起こしテキストに記述されている、長音や促音を含む非言語的特徴を

再現している。長音は長い音として、また促音は短い無音区間として表現される。また、話者は'<u>'タグで示されるが、2人の話者のそれぞれに1人は高い音、もう1人は低い音を割り当て表現する。'vocal'タグで表現される吸気音や呼気音については、単に無音区間として表現している。このプログラムは、それぞれの非言語的特徴を持った単位を繋げることで一連の合成音を作るが、その際音節の連結によるリズム規則にもとづいて単位の長さを変えながら繋げていく[2]。

被験者は、自分の書き起こしテキストからこのようにして作られる合成音とオリジナルの音声とを比較によって、自分の書き起こしテキストの検証と修正を行なった。

表 3: Interpreterble Tags & Entities by Verification Program.

Tags	means
pause	pause
shift	changes of loud,pitch and tempo
u	utterance
vocal	inhale,exhole

Entities	means
&stress;	stress
&trunc;	truncated syllable
&lf;	low fall intonation
&lr;	low rise intonation
&fr;	fall rise intonation
&rf;	rise fall intonation

表 4: Consistency Rate in 5th Transcription[%].

	B-group		W-group	
	%	合計	%	合計
All	75.2%	7,539	81.4%	3,720
Char	91.4%	4,872	91.6%	2,430
Tag	49.4%	2,403	64.2%	1,228
Entity	12.1%	264	25.8%	62

表4に第5セッションの結果を示す。"Char"に関しては91%程度の一致率を得ている。しかし、"Tag"や"Entity"に対しては、"Char"の2/3以下程度の一致率となっている。また、W-groupはB-groupよりも高い一致率を示している。表4の"Tag"は、表2のものよりも高い一致率を得ている。

4 不一致の分析

前節において、タグの一致率は後のセッションになるほど高くなっている。しかしほとんどの場合、タグと実体の一致率は"Char"の約半分程度以下である。図1に、それぞれのセッションにおけるタグの一致率の変化を示す。

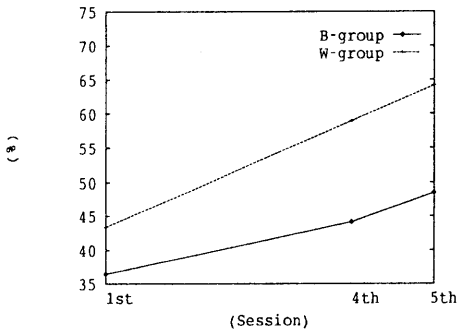

図 1: Development of Consistency Rates of Tags

図1では、それぞれのセッションにおいてW-groupがB-groupよりも高い一致率を得ている。また、第4セッションは第1セッションよりも高い一致率を得ている。ここで、我々はトレーニングとグラフィカル・インターフェースの使用について、第1、第4、第5セッションを通して分散分析を行なった。その結果が表5である。これにより、トレーニングおよびグラフィカル・インターフェースの使用、そしてそれらの交互作用が一致率の改善に効果を及ぼしていることが明らかとなった。

しかし、第4セッションと第5セッションについて、その一致率について検定を行なったところ、第5セッションの一致率は第4セッションよりも高いものの、W-groupのタグを除いて、それが有意な差であるとは認められなかった。

表 6は、 第5セッションにおける一致したタグと実体のについての内訳である。ここで、

表 5: Analysis of Variance of Training and Using Graphical Interface.

	平方和	d.f.	平均平方	F
Training	1140.5	2	570.3	10.3
Using GI	933.3	1	933.3	16.9
Interaction	766.2	2	383.1	6.9
Errors	1160.8	21	55.3	—

"Tag" については上段には会話分析的な情報と言えるものを、そして下段には周辺言語的なものを書いてある。

表 6では両方のグループにおいて、多くの周辺言語的特徴の一致率が、会話分析的なものよりも低くなっている。会話分析的な情報は音声の存在に関連しており、音の存在について被験者は容易に識別できる。結果として、会話分析的情報が周辺言語的特徴よりも高い一致率を得ていると思われる。この差を検定してみたところ、会話分析的情報の一致率の高さは有意であった。

被験者から、非言語的特徴の一致率の低さに関連して以下のような 2 つの問題が報告された。1 つめは、会話をどのように分割するかという問題である。'<u>' タグを例にすれば、'<u>' タグの不一致は被験者がターン・テイキングをどのようにとらえたかによる。'<u>' タグの場合であれば、それほど大きな問題ではない。しかし、他の周辺言語的特徴を現わすタグの場合にも、書き起こしの作業者はそれぞれの周辺言語的特徴の記述に関してのセグメンテーションの基準が必要となる。しかし、作業者がその基準についてそれぞれ異なった認識を持った場合、その書き起こしテキストはそれぞれの作業者によって異なったものになる。この問題は、不明瞭な音や非常に小さい音にも関連していると思われる。仮に作業者がそれらの音を聞きのがせば、その作業者は "<shift feature='loud' new='pp'>" のようなタグを書かないことになる。その結果、その周辺では他の作業者の書き起こしテキストと相違が生じることになる。

被験者から報告されたもう 1 つの問題は、話者のモデルについてであった。書き起こしの作業者は話者の特徴のモデルを構築し、それに従って書き起こしを行なうが、その際に構築したモデルに影響されてしまう。例えば、その話者がしばしば高いピッチで発声する場合を考える。仮に作業者が、その話者は高いピッチで発声するというモデルを構築してしまうと、その作業者はその話者の発声が高いピッチで行なわれているかどうかに充分な注意を払わなくなってしまう。

韻律の書き起こしの場合であれば、作業者間での一致率として約 80% という割合が得られている [4]。我々の実験においては、 非言語的特徴における最も高い一致率であっても 64%(表 4) であった。同一作業者におけるセッション間でのタグの一致率については、第 2 および第 3 セッションでの結果から 80% の一致率を得ている。周辺言語的特徴の分析に際しては、大きなコーパスが必要とされるが、そのような大きなコーパスは 1 人の作業者では構築するのは難しい。そのため、作業者間においてもより高い一致率を得るための手法が必要である。

Alton L. Becker は、彼の短い運動を書き記させるという実験を行なったが、その結果として同じ内容のものは 1 つもなかったと報告している [5]。被験者はそれぞれ異なる視点から彼の行動を観察し、また書き記した。非言語的特徴の書き起こしも、このような動作の記述と同様であろう。今回の実験では TEI に定められた記法を用いた。これは非言語的特徴の記述方法に対して制限を設けることになる。しかし、作業者間での記述の不一致が見られた。これはそれぞれの作業者が互いに異なる視点から音声データに含まれる非言語的特徴を判断し、書き起こしたためであろう。

5 まとめ

本研究では、言語音 (および笑い声など) に対して 91% の書き起こし作業者間の一致率を得た。しかしながら、 非言語的特徴の判断および書き起こしは作業者の主観によって影響を受けるため、 最も高い一致率であってもタグに対して 64% であった。また、作業者の技術およびグラフィカル・インターフェースの使用、合成音を使用しての書き起こしテキストの確認などの一致率への影響も調査した。ここで、作業者の技術およびグラフィカル・インターフェースの使用は作業者間の書き起こしテキストの一致率を向上させるのに有効であった。しかし、合成音による確認については、その効果は明らかにはならなかった。

自然な発話においては、その台本などは存在せず、また不明瞭な音声も含まれる。そのため、

表 6: Inconsistencies of Tags and Entities at Fifth Transcription.

B-group			W-group		
	Total	consistencies		Total	consistencies
Tags			Tags		
Overlap	564	76.6%	Overlap	320	85.0%
Speaker Exchange	243	79.8%	Speaker Exchange	130	87.7%
Pause	159	45.9%	Pause	108	61.1%
Pitch	537	45.8%	Pitch	328	53.0%
Loudness	327	31.2%	Loudness	176	52.3%
Speech Rate	291	34.4%	Speech Rate	116	39.7%
Voice Quality	133	12.0%	Voice Quality	11	18.2%
Nonverbal Sounds	35	57.2%	Nonverbal Sounds	33	66.7%
Others	114	3.5%	Others	6	0.0%
TOTAL	2,403	50.6%	TOTAL	1,228	35.8%
Entities			Entities		
&lr; &lf; &fr; &rf;	57	14.0%	&lr; &lf; &fr; &rf;	22	18.2%
&stress;	108	18.5%	&stress;	22	45.5%
&trunc;	93	4.3%	&trunc;	18	11.1%
Others	6	0.0 %	Others	0	———
TOTAL	264	12.1%	TOTAL	62	25.8%
TOTAL	2,667	45.7%	TOTAL	1,290	62.3%

本実験で得られた言語音に対しての作業者間での一致率はかなり高いものと言える。しかし、タグおよび実体に関しては、それらの作業者間での一致率は言語音のおよそ半分以下であった。非言語的特徴に関しては、それぞれの作業者が互いに異なる視点から特徴を判断するためである。分析のためのコーパス作成という点からは、可能なかぎり作業者間での一致率が高くなるような表記方法や基準が必要とされる。

参考文献

[1] George L. Trager: "Paralanguage: A First Approximation", Studies in Linguistics, 13, 1-11, 1958.

[2] 加藤 雅代, 古村 光夫, 橋本 新一郎: "母音部エネルギー重心点に着目した日本語リズム規則", 音響誌, 50, 11, pp. 888-896,(1994-11).

[3] C.M.Sperberg-McQueen, Lou Burnard: "Base Tag Set for Transcription of Spoken Texts", TEI P3 chapter 11, Text Encoding Initiative, Chicago, Oxford, 1994.

[4] Kim Silverman, Mary Beckman, John Pirelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, Julia Hirschberg: "TOBI: A Standard for Labeling English Prosody", Proc. of ICSLP 1992 pp. 867–870.

[5] Alton L. Becker, "Language in paricular: A lecture.", In D. Tannen (Ed.), Liguistics in context: Connecting observation and understanding, Ablex pub. corp., 1988.