書き起こし困難点の作業コストを低減する支援ツールの設計

Design for a support tool to reduce the coding cost for spoken multi-party discourse

東新　順一　　竹内　和広　　水上　悦雄　　森本　郁代

Junichi Toshin　　Kazuhiro Takeuchi　　Etuo Mizukami　　Ikuyo Morimoto

大阪電気通信大学
Osaka Electro-Communication University
国際電気通信基礎技術研究所
Advanced Telecommunications Research Institute International
関西学院大学
Kwansei Gakuin University

Abstract: Transcription of spoken discourse plays a central role in multi-party discourse analysis. Previously, transcriptions have usually been coded by experts, but recent developments in speech technology make it possible to process signal data automatically. In this paper, we propose a design for a tool which displays a complicated signal as a graphical representation using various signal processing modules. The advantage of this tool is to be able to reduce troublesome phenomenon in coding para-linguistic signals occurring from non verbal elements.

1　はじめに

近年，会議議事録自動作成などを目的とした多人数対話処理システムに対する期待が高まっており，様々な観点や方法による多人数インタラクション研究が国内外で行われている[1]．会議分析に用いられるデータにおいても様々な工夫がみられ，必要に応じて収録方法や生成方法の効率化が検討され質のよいデータが求められている[2].

また音声の抑揚や強弱，速さ，声質といった情報は，音声言語コミュニケーションにおいて重要な役割を果たすことが知られている．例えば，韻律情報を用いた発話の感情分類や，肯定的発話か否定的発話かを同定する研究が行われている[3]．つまり，会話構造を記録するためには文字情報だけでは不十分であり，従来文字化されてこなかった音声情報をいかに分析し記述するかについての課題も重要である．そのような情報には，男性女性の話者の性別といった話者属性の区別や，感情や心的態度といった情報の伝達や会議の状態とも密接な関係がある情報も含まれる．

音声言語処理の研究では，既に，このような情報を研究資源として事後に利用しやすい形で記述・表現するためのコーディング手法が検討されており[4]，イントネーション情報の記述・表現モデルは数多く提案されている．例えば，日本語においては ToBI を日本語に適用した J-ToBI の有効性が示されている[5].

J-ToBI のラベリングが人間の韻律構造の分析記述を目的としていることに対し，本稿では，会議などの会話を映像や音声により記録し保存・利用する機会が増えてきていることを背景として，会議を分析する上で必要となる書き起こし作業を支援するツールについて検討する．具体的には会議などの多人数会話における書き起こしについて，分析者の要求や作業の困難点を提示した上で，作業コストを低減するツールの設計を行いたい.

2　多人数会話の書き起こし

2.1　人手による書き起こしの必要性

実データを元に分析を行う対話研究においては，音声，映像データなどの１次データに加えて，音声を文字に起こした, 書き起こしデータが必要になる．会話データのコーパス化を目指す場合は言うまでもないが, コーパスを作成することを目的とせずとも，

非言語情報やパラ言語情報のみを対象とするような研究アプローチでもなければ，書き起こしデータがあることが，分析を進める上での前提となる．例えば，会話分析においては，対象とする会話データを文字と独特の記号を用いたトランスクリプト（書き起こし）の作成[6]は，研究を進める上でも，発見した事実を誌上で説明する上でも，ほぼ必須の作業となっている．

かつ，書き起こしデータは，音声自動認識の技術が格段に進んでいる現在にあっても，人手によって作成されているのが現状であり，そもそも音声自動認識技術の開発そのものに，正解データ，学習データとしての人手で作成されたコーパスが必要となる．特に，複数人数が関わる多人数会話のデータであれば，現実問題として，人手に頼らざるを得ない．

2.2　多人数会話の書き起こしの現状

音声の書き起こし作業は，現在ではPC上に音声ファイルや動画ファイルを取り込んで，音声分析あるいは書き起こし支援ソフトウェアを使用して行うのが一般的である．その際，多人数ならではの問題の第一は，音声の話者分離である．よほどの設備を整えて収録するのでもない限り，通常は全員の音声が混ざったモノラルもしくはステレオの音声ファイルから話者を特定する必要がある．話者分離技術に関しては，既に多くの研究が進められており，有効性が示されている[7]．本研究でも，将来的には話者分離技術を実装することも射程にいれてはいるが，現状では，話者ごとに分離されたマルチチャンネルの音声ファイルを扱うことを前提としている．これを前提として，3 チャンネル以上のマルチチャンネル音声ファイルに対応したソフトウェアとして現在利用可能なものに，Wavesurfer[1][8]，Multitrans[2][9]がある．例えばMultitransは，図 1 に示すように，書き起こし領域と音声波形表示部がわかれていて，かつそれらが相互参照可能になっている点，選択した話者の音声のみが再生可能である点，発話区間の分離・統合が容易である点など，書き起こしの作成を志向した，数少ない支援ソフトである．

しかしながら，例えばノート PC で作業をする場合には，話者が増えるとディスプレイ画面に入りきらないという問題や，このことに関連して，波形の

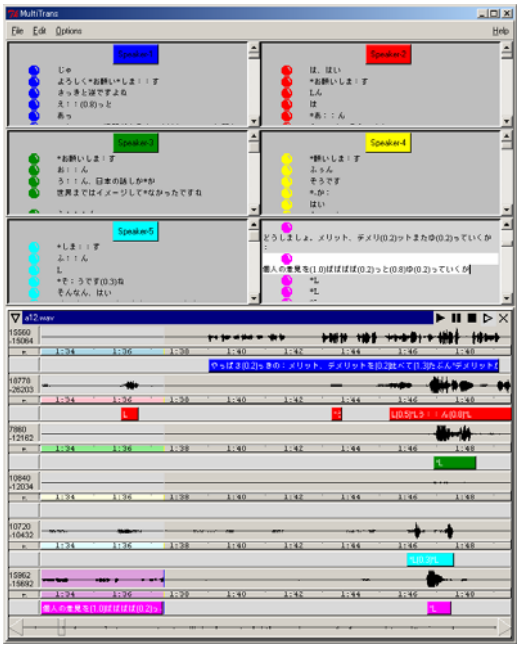

図 1　Multitrans の作業画面

拡大や，これ以上のラベル行の増加は画面が煩雑になって見にくいという問題など，実作業上，不便な点がいくつかある．

また，多人数が関わる自然な会話の場合には，その話者一人の音声だけでは内容が判別できないような発話も，別の話者の先行発話と一緒に聞くことで判別可能になる場合があり，特に，三人以上の複数人が入れ代わり立ち代わり発話をするようなケースでは，各話者の音声を何度も確認する必要がある．逆に精密な発話の書き起こしを行うために，話者一人の音声を一時的に聞きたい場合もあり，切り替えの柔軟さが必要である．さらに，次節で説明するように，オーバーラップ箇所を特定しようとした場合の手間は相当なものであり，単純に，話者一人一人の音声を人数分だけ書き起こすという作業には終わらず，多人数会話に特化した書き起こし支援ツールが求められる．

3.書き起こしの困難点

3.1　書き起こしの難しさ

書き起こしは，単純に発話された音声の言語情報だけを書き起こせばよいわけではない．書き起こしデータから，会話のやり取りを理解するための，最低限の韻律的な特徴や，相互行為上の要素が，含まれていることが望まれる．例えば，図 2 に示すように，言語情報を表現した文字だけではなく，言語に付随する情報を示した記号が付与される．

1　ただし，著者らの知る限りにおいてver. 1.6 以前のバージョンでしか，マルチチャンネル音声の正常動作を確認できていない．

2　5 チャンネル以上の音声ファイル，および日本語への対応は，千葉大学伝康晴氏の作成したパッチプログラムによる．使い方等は，小磯(2006)[10]も参照のこと

{ゼミ合宿プランの打ち合わせ}

11.07.11　11.14.24 C:ええっと(0.2)お風呂：なんですけれども：(0.6)お風呂は温泉じゃなきゃ：とかありますか①？
11.15.45　11.19.36 B:ん②：：：どうせなら(0.4)温泉の方が(0.3)いいかな：っと③[思います　]
11.18.67　11.19.89 A:　　[うん：：：]
11.20.25　11.23.36 B:④hh[hh]
11.21.73　11.25.41 A:　[hhh]ですよね：h[hhh　]
11.24.17　11.27.25 C:　　[hh やっ]ぱり：(0.3)そういうのってありますか：
11.28.77　11.30.38 A:温泉気持ちいいですもんね：：
11.30.90　11.31.98 B:あっ⑤(0.2)でもお部屋にも[：：]
11.31.52　11.32.27 C:　　[はい]：
11.32.36　11.33.47 B:あったらいいな：っと
11.33.55　11.34.16 C:はい[：：]
11.33.78　11.34.21 B:　[ん：]
11.34.63　11.35.89 C:わっかりました：

図 2　書き起こしデータ例

以下にはツールによる支援が有益である可能性が高い，書き起こしが難しい情報についての説明を行う．

【韻律的要素】

発話が上昇調で終わっているのかなどのイントネーションを示す記号（例えば“？”図 2 中①）や，語末の伸張を示す長音記号（例えば“:”図 2 中②），発話と発話の間のポーズ長などは，どのように発話がなされていたかを書き起こしから推し量るために有用な情報となる．著者らのこれまでの経験上，例えば，語末のイントネーションが上昇調であると判断するかどうかは，その書き起こし作業者に依存する．また，長音記号を付与する際に，どれだけ伸張されたかに対応して，長音記号を複数付与するが，この長さの基準は「拍（モーラ拍）」とされることが多く，これもやはり，書き起こし作業者間でゆれが生じる．

【やり取りに関わる要素】

各発話が開始された時間情報は機械的に抽出可能でも，現行話者がどの音を発しているときに，他者の言い重なりが始まったかという情報は音声情報を参照しなければならない．オーバーラップ地点を示す記号（例えば“[”図 2 中③）は書き起こしだけからこれを判断するために必要な情報となる．特に多人数会話特有の問題として，三人以上が代わる代わる言い重なっているような箇所は，誰が誰に言い重なったのかの対応付けは容易ではない．

【ことば以外の音声要素】

自然な会話においては，笑いは少なからず生起するし，笑いながらしゃべることも頻繁に見られる（例えば“h”図 2 中④）．笑いがやり取りにおいて重要な意味を持つことは言うまでもないが，これを書き起こすことは単純ではない．その他，咳払い，ため息，空気すすりなど，言葉としてそのまま書き起こすことが難しい音声要素も，相互行為上，少なからず意味を持つようなものは，無視はできない．しかし，マイク性能に依存して，通常は聞こえないような鼻息や息継ぎの音が入っていた場合に，笑いや空気すすりとの判別が困難な場合がある．さらに，笑いに関して言えば，どこまでを笑いとするのか，笑いながらの発話はどう表記すればよいのかの基準化は容易ではなく，書き起こし作業者によって，ゆれが生じざるを得ない．笑いを分析対象とするような研究であれば，この問題はより深刻なものとなる．

3.2 書き起こし支援ツールに期待されること

3.1 節に挙げた点の多くは，分析者が必要とする（無視できない）要素例であると同時に，その基準化が困難なものでもある．基準化が困難であることは，すなわち書き起こし作業者間でのゆれにつながる．

もし，音声の韻律的特徴量等を用いて，前節に挙げたような要素の基準化を支援するような仕組みが実装されていた場合には，作業者間のゆれは大幅に減少するであろう．

また，最も重要なことは，分析者の要望に合わせて，情報の要／不要がカスタマイズ可能であることであろう．さらに言うならば，書き起こし作業の最終目標は，入力ではなく，その出力にあるため，書き起こしとして出力される形が，分析者の望む形となっていることが期待される．分析者によっては，一定の基準で区切られた発話区間が，そのまま時間順にソートされて出力されることを望む者もいれば（Multitransはこの形（lcf形式）で出力する），一定長以上のポーズで区切った発話区間が一つの発話文としてつながっている場合（つまり発話内ポーズであれば），その間のポーズ長を括弧内に記述した上でつなげて出力し(図 2 中⑤)，さらにはオーバーラップ地点が揃うように出力されることを望む分析者もいるであろう．この選択のバリエーションがあるかないか，あるいは，ソフトウェアの枠組みとしてその拡張可能性があるかどうかは重要であろう．

4　書き起こし支援ツールの設計

前節で述べた多人数対話の問題に対応するための書き起こし支援ツールの設計を行った．このツールは人手での作業の効率化や書き起こしにおける作業者の基準の一貫性をツールによって支援することを目的としている．

4.1　書き起こしを支援する音声信号表示

今回のツール設計の基本的アイディアは通常音声信号を図 1 の Multitrans のように波形として表現するのではなく，音声信号をグラフィカルに表示する点である．また，音声信号に対していくつかの信号処理を行い，認定に専門性が必要な「笑い」や「空気すすり」といった部分も表示をする．例えば，音声の特定部分が「笑い」の音声信号含んでいる部分は，「笑い」部分を自動認識する技術を援用し，時間単位ごとの「笑い」らしさをグラフィックで表示する．

具体的に説明すると，一般的な音声信号は横軸を時間位置，縦軸を音の大きさとして音声波形表示をする．それに対し，音の大きさを色の濃さに対応させ，音の大きい部分は色を濃く，小さい部分は薄くというように，色を使ってグラフィック表示ができる．この対応を図 3 に示す．このことによって，横軸が時間位置になる関係は崩さずに，縦方向にスペースを小さく表示が可能となる．これが基本的なアイディアである．

図 3　音声信号とグラフィックの対応

さらにこれに，単に色の濃淡を音声の強さに対応させるのではなく，信号処理の技術を使って，例えば「発話」らしさ，「笑い」らしさや「発話末の伸び」を数値化し，それを図 3 のようにグラフィックに変換し，同じ横軸上に表示させたイメージが図 4 である．

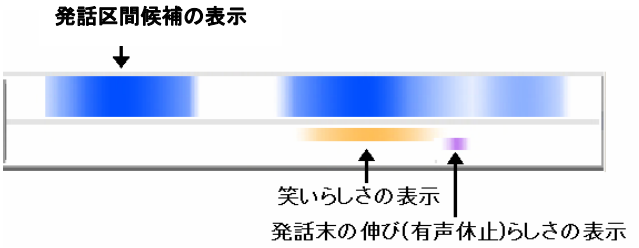

図 4　音声情報の表記方法

以上のような一話者における複数の音声信号特徴を時間軸上にグラフィカルに表示させるアイディアを使った，多人数の書き起こし支援ツールのイメージを図 5 に示す．

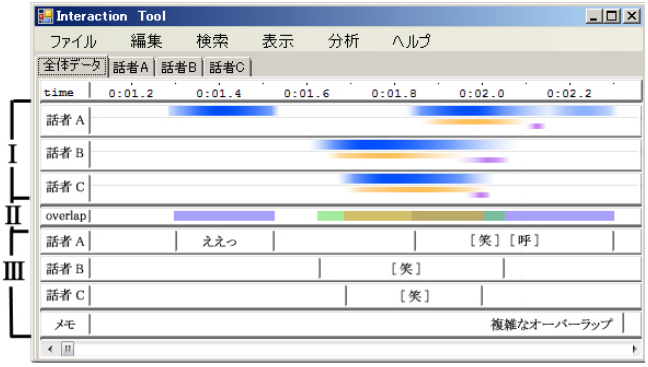

図 5　書き起こし支援ツールのイメージ

図 5 の I の部分は，上で説明した時間軸に対して音声信号をグラフィック表示したものを複数人分表示している部分である．

II の部分は，その複数人分の音声信号の，「発話」が重複している(オーバーラップ)箇所を，音声信号と同様にグラフィック表示する部分である．2.2 節で述べたように，会話参加者の発話の相互関係は，多人数会話書き起こしを難しくする要因となっている．

それに対して，我々の設計における表示では，一目で作業者がオーバーラップ部分や話者構造を特定することに寄与する．例えば，発話オーバーラップが起こっている場合，色の重ね合わせを利用して色彩を変化させる等の工夫が考えられる．また，「発話」だけの相互関係を表示する必要はなく，例えば，全員が同時に発話している場面(図 5 においては笑っている場面)を他の部分と色分けて表示することも考えられる．

Ⅲの部分に関しては他の書き起こしツールと同様，発話情報を書き込むことができ，さらに支援機能を付随させる．書き起こしを行う際に作業コストが大きくなる作業のひとつに発話区間を特定することがあげられる．そこで図 6 のように自動的に仮の発話区間を示し，作業者に仮発話区間を修正してもらう形をとることで，作業コストの低減を実現する半自動の認定機能を考えている．

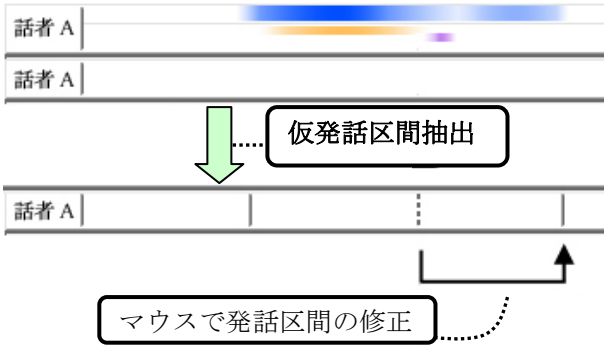

図 6　発話区間特定の様子

ツールの作業画面は，作業者のカスタマイズにより音声波形や基本周波数などさまざまな情報の表示・非表示，配置の設定は自由に変更できるようにしたい．これは音声波形による音声波形による書き起こしに慣れている作業者への対応や，書き起こし内容により必要な情報が異なるためである．

4.2 各特長の解析モジュール

音声信号から「発話」や「笑い」といった特徴表示を行う候補としては，以下のようなものを考えている．もちろん，その他の特徴の追加も含めて考えている．

A1.あいづち　　A2.空気すすり
A3.有声休止　　A4.咳払い
A5.ため息

このような発話に関連する音声特徴を数値化するモジュールを作成する．

現在，我々のグループでは，以上のような特徴解析モジュールとして発話区間と「笑い」らしさを解析するモジュールから開発を進めている．

具体的には，発話区間の解析は，収録音声の広い部分から発話者の音声特徴と録音状況の特性を求め，当該話者の音の強さの傾向と周波数特性を利用して，当該話者の発話区間を推定する方法の開発を進めている．また，音声データの中に混在するスクラッチノイズなどのノイズ除去も同時に行っている．

また，「笑い」らしさの特徴解析には，上記で求めた発話区間前後を分析対象とし，Wavelet 解析を用いた特徴抽出，ニューラルネットワークを用いた機械学習による特徴分類といったアプローチで研究を行っている．

図 7　出力設定画面

4.3　付随支援機能

2.1 節で説明した通り，書き起こしは後のデータ分析に使うためのデータ作りであるため，書き起こし作業の効率化を図ると同時に，書き起こされたデータを使いやすい形に整形して出力する必要がある．

例えば会話文を単なるテキストとしてそのまま出力しただけでは分析を行う際に扱いにくく，適切にスペースや記号を入れて整形することが必要である．特にオーバーラップに関してはどの記号が対をなしているのかがわかりにくい．そこで記号の対応付けを自動で行う機能を設計した．また 3.2 節でも述べたが，ファイルの出力関しては，分析単位をどう定義するかで表記方法が変わってくる．分析単位は分析者により異なるため，出力に関しては細かい設計を行う必要があるといえる．図 7 に，現在設計段階

にある出力の設定画面を示す．発話やポーズの表記の仕方，書き起こしの際の記号をどう整形して出力を行うかが，選択するだけで変更できる．

もう一つの機能として，書き起こし前に事前に簡単な分析を必要とする場合や特定現象部分を確認したい場合に対応するために，データの統計・検索機能を用意して設計した．特定現象の持続時間や回数から頻度，総計を算出する，また持続時間や値の大きさ，ラベリング情報から該当部分へ移動を行うといった状況を想定し，さらに閾値による絞込みも考慮している．例えば，「笑い発話が全体の何%を占めている」や「発話重複部分が何回あった」という統計情報が出せれば分析を行う際に非常に役立つと考えられる．この機能は分析者が全ての発話に関してではなく，分析が必要な特定の特徴(例えばオーバーラップが起こっている)の部分のみを書き起こし，作業を行うことに役立つのではないか．

5　おわりに

本稿では書き起こしの必要性と問題点をあげ，その問題に対応できる書き起こし支援ツールの設計について述べた．設計したツールの特徴は「笑い」や「空気すすり」といった，発話の様態に係る音声特徴をグラフィカルに提示することである．また，その表示のための機構を利用し，書き起こし作業者間におけるゆれを抑える支援機能を盛り込む．さらに書き起こしたデータの出力部分も柔軟に改変できる機能も加えた．このように，本稿が提案するツールは，実際の書き起こし作業者，書き起こしデータを利用する側の利便性を重視した設計となっている．

現在，我々のグループでは，この設計での書き起こし支援ツールの実装を行っており，書き起こし作業のコスト削減と書き起こしデータの利用に必要なデータ整形のコストを削減する検証を行っている．今後，設計した機能の実装をさらに進め，実用的なツールとして完成させていきたい．

謝辞

本研究は，（独）科学技術振興機構・社会技術研究開発センター研究開発プログラム「21 世紀の科学技術リテラシー」平成 19 年度採択課題『自律型対話プログラムによる科学技術リテラシーの育成（研究代表者：大塚裕子）の助成を受けて行われた．

参考文献

[1] 坊農真弓，高梨克也．連載チュートリアル「多人数インタラクション研究には何が必要か？」〔第１回〕インタラクション研究の国内外の動向と現状．人工知能学会誌，Vol.22，No.5，pp.703-709，2007.

[2] 水上悦雄，森本郁代，鈴木佳奈，大塚裕子，竹内和広，東新順一，奥村学，柏岡秀紀．話し合いにおけるコミュニケーションプロセスの評価法について．第 14 回言語処理学会年次大会発表論文集，2008.

[3] 藤江真也．パラ言語情報を利用した音声対話システムに関する研究．早稲田大学修士論文 2005.

[4] 菊池英明．連載チュートリアル「多人数インタラクションの分析方法」〔第 3 回〕音声言語コミュニケーション研究のための分析単位 -ToBI- 人工知能学会誌，Vol.23，No.1，pp.113-118，2008.

[5] ニック キャンベル．解説 Tones and Break Indices(ToBI)システムと日本語への運用．日本音響学会，Vol.53，No.3，pp.223-229，1997.

[6] 好井裕明，山田富秋，西阪仰(編)．会話分析への招待．世界思想社，1999.

[7] 秋田祐哉，川原達也．多数話者モデルを用いた討論音声の教師なし話者インデキシング．電子情報通信学会論文誌，Vol.J87-D-II，No.2，pp.495-503，2004.

[8] Wavesurfer
http://www.speech.kth.se/wavesurfer/

[9] Multitrans
http://agtk.sourceforge.net/

[10] 小磯花絵，（伝康晴，田中ゆかり編）．講座社会言語科学６－方法，会話データの構築法．ひつじ書房，pp. 170-186，2006.