IE3-02

# 韻律情報に基づく映像編集支援システム

## A video support system based on prosodic information

永田 絢子[*1] 桐山 伸也[*2] 北澤 茂義[*2]
Ayako Nagata Shinya Kiriyama Shigeyoshi Kitazawa

[*1] 静岡大学大学院情報学研究科
Graduate school of Information, Shizuoka University

[*2] 静岡大学情報学部
Faculty of Information, Shizuoka University

A topic turning points and unnecessary part's becoming the clue of the edit, and presenting them lead to the reduction of shortening the edit working hours and the work load. From the viewpoints, the automatic detection technique of the topic turning points where pause position was made a candidate was examined based on prosodic information which aimed at the construction of the a video edit support system.

## 1. はじめに

近年 Web コンテンツ等, 映像データを目にする機会が増えている. しかしデータ量は膨大でその中の多くが自然発話によって会話が行われるため, 不要箇所が多く含まれ編集作業には大変な労力と時間がかかる. また, 編集の際データを最初から最後まで見なければいけないという点も作業負担, 作業時間を増やしている. そこで重要部分を自動検出し, 自動編集が出来るシステムがあれば大変便利である. しかし, 編集はどこを使いたいのか, 何を伝えたいかにより出来上がりは異なり, それは編集の意図を反映させたものでなければならず, その仕組みが必要になる. そこでまずは編集作業を支援する情報について考える. 例えば話題転換部を表示することで, どこで何を言っているかが話題の冒頭部分のみを見れば分かり, 時間短縮につながる. また転換部間の発話の多さにより盛り上がっている話題か, そうでないか,やキーワード入力によりその話題がどこで話されているかを検出できるようになると考えられる. また不要箇所を提示することでそこを見飛ばすことが出来たり, あらかじめ自動カットしたものを提示することが可能になると考えられる.

今回はまず, そのための第一歩として音声データに着目し, F0 やパワーといった韻律情報とキーワードに着目し, 話題転換点や不要箇所を提示することによって映像編集作業の支援を行う.

## 2. 話題転換点の特徴

映像編集を行う際, 編集者は映像を見てどの話題を使うか決めていく. そこで話題の転換点を提示することは非常に役に立つ.

話題転換点はどのようなデータを扱うかによってその特徴は変わる. 例えば講演音声では「次」や「結果」, 「研究」, 「実験」, 「評価」等の単語が談話標識となることから,それらの談話指標に基づく重要文抽出が行われている[1].

今回は特にインタビュー音声に着目し, 特徴分析を行った. 使用したのはいずれも 1 対 1 によるインタビュー形式で概要を表 1 に示す.

### 2.1 インタビュー音声

インタビューではインタビュアーからの質問に対し, インタビュー を受ける側がそれに答えるといった形が繰り返される. そこでのインタビュアーの役割は「話題を聞き出すこと」と「話題を広げること」である. これは次の話題への転換, 話題の詳細への転換に当たる. これによりインタビューは進んでいくことからインタビュアーの発話が話の流れの鍵を握り, それにより話題転換が行われていると考えられる.

表 1.使用データ概要(48000Hz 16bit mono)

| | 話者 | インタビュアー | 時間(min) | 転換点 | 割合 |
|---|---|---|---|---|---|
| ff01 | 女性 | 女性 | 8.59 | 27 | 0.96 |
| ff02 | 女性 | 女性 | 4.48 | 15 | 0.93 |
| ff03 | 女性 | 女性 | 6.11 | 21 | 1.00 |
| ff04 | 女性 | 女性 | 9.16 | 26 | 0.92 |
| mf05 | 男性 | 女性 | 33.44 | 35 | 0.74 |
| mf06 | 男性 | 女性 | 22.32 | 42 | 0.26 |
| mm07 | 男性 | 男性 | 11.29 | 27 | 0.89 |

そこで話題転換点と話者交替について関係を調べるため, 実際に音声データに話題転換点と話者交替点へのラベリングを行い, 調査を行った. その結果, 7 名中 6 名の話者で話題転換の 7 割以上が話者からインタビュアーへの話者交替点, またはその後のインタビュアーの発言中に話題転換が行われることがわかった(結果を表 1－割合に示す). 特にその割合が低かった mf06 は特徴的に一つのコメントが長いことにより, その中にいくつも話題が上ることでその割合は低くなった. しかし, その転換点にも特徴が見られ「(それ)で」, 「したがって」, 「だから」等の接続詞による転換が多く見られた.

### 2.2 インタビュアーの発話の特徴

・インタビュアーの発言は質問や問いかけであることが多いため, 発話末に「～ですか?」, 「～ますか?」, 「～ですね?」, 「～(です)よね?」といった特徴的な発話があらわれる. またそれに伴い語尾が上がる現象が見られる. しかし, 相手の話者によってはそれを言い終わる前に発言を始めてしまうため,その部分がかき消されてしまう場合もある.

・発話の冒頭の特徴としては, 相手の発話を受けて「分かりました」, 「なるほど」, 「そうですね」といった相槌が現れる.

連絡先:永田 絢子 静岡大学大学院

### 2.3 インタビューを受ける話者の特徴

・インタビュアーの質問を受けて,「はい」,「そうですねぇ」といった相槌が一発話の先頭で見られる. この特徴は ff01～ff04 までに共通してみられた.
・インタビューが技術などの説明を求める内容であると冒頭には「それは」,「これは」,「あれは」といった発言がみられ, mm07 のケースでは 9 割以上の発話冒頭部で観察できた.
・コメントが長くなっているときは自分で話題を転換している傾向にある. その多くで「(それ)で」,「だから」,「従って」といった話者特有の接続詞がみられる.

## 3. 話題転換点抽出法の検討

第 2 節で触れたとおり, 話題転換点, 話者転換点にはいくつかの特徴が見られた. その特徴を利用し,話題転換点の抽出法を検討していく.

まず一つ目に考えられるのがワードスポッティングにより話題転換点に特徴的語を抽出する方法である. うまく抽出することが出来れば, 話者交替点, 話題転換点の検出が行える. しかしこれは音声認識結果の影響を大きくうけるため, 認識結果のいいものでないと使えないという問題点がある.

次に考えられるのがポーズを検出することで話題転換点の候補点を抽出する方法である. これは話者交替が発話の区切りで行われるためにポーズを含むのではないか, また, 話者転換点以外での部分でも一呼吸置いて次の話題に移っていくのではないか, という予想に基づくものである.

そこで,まず始めにポーズ位置の検出を行い, それを候補とした話題転換点の抽出法を検討した.

### 3.1 ポーズ位置を利用した話題転換点候補の抽出

まずは話題転換点候補としてポーズの検出を行う.

使用したデータは 2 節で紹介したインタビュー音声のうち ff01 を使用し, 音声データに手動でポーズ位置のラベリングを行い,パワーの平均値を求めた.詳細を表 2 に示す.

表 2.ff01 音声データ

| ラベル | ラベル数 | パワー平均(dB) |
|---|---|---|
| 発話 | 200 | 43.57 |
| ポーズ | 202 | 29.75 |

#### パワーによるポーズ検出

ポーズのパワーの平均 29.75dB よりそれぞれ閾値を 31dB に設定し, 閾値以下の区間が一定以上続く区間の検出を行った. 今回は発話間のポーズを検出できればよいため時間の閾値は 150msec とした. またこれだけではポーズ中にノイズがのることで連続するポーズ区間が分断されてしまうため, ポーズではさまれる 100msec 以下の区間はノイズとみなし連続するポーズに置き換えるという処理を行った.

手動で付与したラベルを正解ラベルとし, 再現率は正解ラベルのラベルに対しどれだけ抽出できたか, 適合率は抽出できたものの内正解(手動ラベル区間中を抽出)したものの数により算出した. 正解の判定については, 今回は転換点として検出できれば良いので区間の長さは気にせず, 正解区間中を抽出できれば正解とみなした(ずれは 20mse まで許容). また, 一つの正解区間に複数の抽出区間のある場合は正解数を 1 とした.

結果は再現率 0.72, 適合率 0.67 という結果になった. この方法により, 十分な話題転換候補点を検出できるかを調べるため, 別の音声ファイル(ff02～mm07)のポーズ検出も行った. 閾値は

表 3.候補点概要

| データ | 抽出数 | 転換点数 | 抽出数 | 適合率 | 再現率 |
|---|---|---|---|---|---|
| ff01 | 139 | 27 | 19 | 0.137 | 0.704 |
| ff02 | 124 | 15 | 13 | 0.105 | 0.867 |
| ff03 | 62 | 21 | 17 | 0.274 | 0.809 |
| ff04 | 292 | 26 | 23 | 0.079 | 0.884 |
| mf05 | 796 | 42 | 38 | 0.029 | 0.905 |
| mf06 | 389 | 35 | 29 | 0.075 | 0.829 |
| mm07 | 420 | 27 | 27 | 0.065 | 1 |

データにより背景の雑音レベル, 話者の声の大きさが違うため, その時々で変更するのが好ましい. しかし, 実際のデータにはラベルはついておらず, ポーズ区間のパワーの平均を知ることは難しい. そこで今回は各音声データの開始 30 秒分の無音区間, 発話区間を手動でラベル付けをし, そのポーズ区間の平均パワーからそれぞれ個別の閾値を設定した. それぞれの抽出数とそのうちの転換点のポーズ検出数,転換点の検出数を表 3 にまとめる.

今回のポーズ検出は話題転換候補点の検出を目的に起こったので, どれだけ転換点部分を抽出できたかが重要になる.

話題転換点に着目してみてみると転換点そのものが検出出来た割合は約 7 割となった. これは相手の発話(相槌)が被ってしまう, 息を吸い込む音, などといった原因でポーズとして検出できないことによるものである.ここでは全ての転換点を候補としてあげる必要があるので, ポーズ検出性能の向上が求められる.

データについてみてみると, 相槌を打つことなどにより話者交替箇所とずれた位置で起こる話題転換点も候補として抽出できるという利点もあった.

またもともとポーズではない点で話題転換がなされる場合もあるので,別の方法(韻律的特徴による話者交替の検出等)で候補点として挙げていくことを考えなければならない.

一方, 抽出してきたポーズに対する話題転換点の割合は 10%程度(高くても 27%)なので, さらに制約をかけて絞っていく必要がある.

### 3.2 候補点からの話題転換点の抽出の検討

話題転換部抽出のためにはこの候補点の中から, さらに制約を加え実際の転換点を抽出する必要がある. そこで話者交替点ではポーズの前後で音声の特徴が変わることに着目し, さらに候補点を絞り話題転換点の抽出を行う.

今回は話者毎に声の高が違う点に着目し, ポーズの前後で「F0 の平均が一定以上変わる」という条件を加え, 候補点からの転換点の抽出を行った. 方法は 100msec 毎の F0 平均をポーズとポーズではさまれる区間について調べ, ポーズの前後の区間平均の誤差が閾値以上になる点を話題転換点とした.

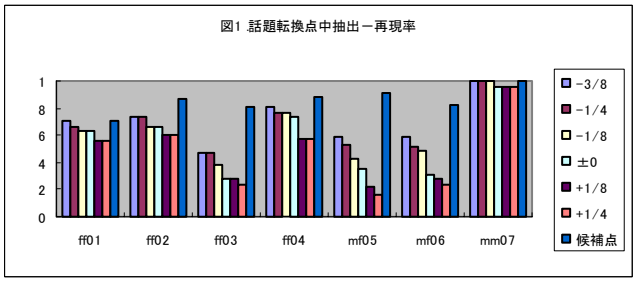
閾値決定のために予めデータの冒頭 30 秒間の各話者の平均 F0 を求め, インタビュアーと受ける側の F0 平均の差を求めた.結果を表 4 に示す.

表 4.各データの話者間の平均 F0 の差

| ff01 | ff02 | ff03 | ff04 | fm05 | fm06 | mm07 |
|---|---|---|---|---|---|---|
| 23.87 | 16.34 | 19.01 | 30.79 | 124.50 | 117.52 | 14.38 |

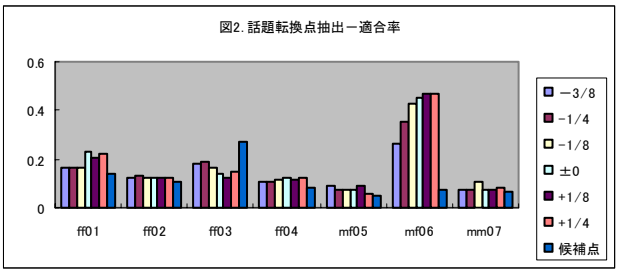
表からも分かるとおり, 話者間の F0 平均の差はばらばらでデータにより開きがある. 実際のシステムを考えたときには自動的

に閾値を決定しなければならない. そこで今回は求めた F0 平均の差から相対的に閾値の決定を行った. 用意した閾値は F0 平均の差の値を 8 分し, 8 分の 3 を引いたものから 8 分の 2 を足したものまでを閾値とし, それぞれ転換点抽出を行った.結果を図 1,2 に示す.




**再現率**

再現率は ff01, mm07 以外で候補点からの抽出を行ったことにより, 大きく下がった. 再現率が低くなった理由としては,候補点に絞る時点で転換点が脱落してしまったことが挙げられる.特にそれが多かったのが ff03, mf05 で, ff03 では二話者の声の高さの差がそれほどなかったこと, mf05 では 1 つのポーズとポーズの間の区間で話者転換が行われる箇所があり, F0 の平均が二話者の発話を合わせたものになってしまったため, 前の区間との差が現れず, 転換点として抽出できなかった. この問題には終わりは質問文などの語尾のトーン変化の影響を受けやすいため,始めの一定区間のみの F0 を使う, などといった対応策が挙げられる. また, ff01, mm07 では閾値による変化は見られなかった.




**適合率**

適合率は mf06 以外の話者で候補点と大差ない結果におわった. この理由として, F0 は話者内変動を受けやすく同一話者間でもポーズの前後で F0 の差が大きいことが多いため, 転換点を絞りきることが出来なかったことがあげられる. また mf06 はインタビュアーと受ける側の性別が異なるため, F0 による閾値を設けることで, 多くの候補点を対象からはずすことが出来き, 他のものよりいい結果になった. このことから F0 による判定は話者の性別が異なるときに有効なパラメータになり得るといえる.

多くの話者で話者交替部と転換点のかかわりが深いことを考えれば, 話者交替点としての絞込み部分での脱落は防ぐべきである. そこで今回は声の特徴として F0 を使用して抽出を行ったが,F0 だけでは話者交替を識別するのに不十分であるので, スペクトル包絡等のパラメータを加えることで, 話者交替点の抽出を行う必要がある. F0 平均の差に関しての閾値は低めに設けることで転換点の脱落を防ぎ, 候補を減らすことに利用できる.

また話者交替点以外(インタビュアーの発話中の転換, 接続詞による転換)での候補点を絞り, さらには脱落させないために, ポーズの前に相槌が打たれる, ポーズの直後に接続詞が現れるといった特徴も並行して用いた抽出法の検討が必要である.

## 4. システム設計

今回の検出法では十分な精度は得られなかったものの, より正確に抽出することで話題転換点は編集作業に非常に役に立つ情報となる. そこで話題転換点を利用した映像編集支援システムの構築を検討する. 全体のシステムの流れを図 3 に示す.

**重要部抽出**

今回抽出した話題転換点をもとに, 転換点までを一つの単位とし, その中から重要話題の自動抽出を行う. しかし,編集の意図によりどこが重要部になるかは変わってくる. そこで編集者の意図を重要部判定にうまく反映させる仕組みが必要になってくる. 例えば, キーワードを入力し, そのキーワードが多く現れる部分, 転換点間の発話量を調べ盛り上がっている部分などを重要部とする方法が考えられるが, 重要部の定義についても検討が必要である.

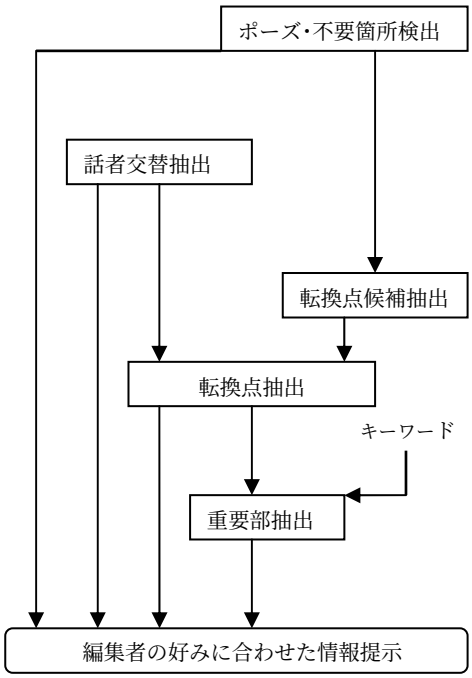
**提示情報と提示法**

話題転換点,重要部のほかに, 不要箇所(ポーズや言い直し, 言い淀みなど)の抽出を行い, 編集に役立てることを考える. これらを自動抽出できればあらかじめ不要箇所をカットした映像を提示することで,映像を見る時間を短縮できるようになる.

また実際の編集作業では映像のつながりも重要になってくる. そのため, ポーズ位置を提示することで,より自然な場所で映像を区切ることが出来るようになると考えられる. 他にも話者交替を正確に抽出することにより, どの話者が発話しているかを提示できるようになり, インタビュアーのみ, インタビューを受ける側のみの発話が聞けるようになる.

実際のシステムとして編集者が使うときには, 提示法についても重要になってくる. 情報提示が編集の妨げにならないよう編集者の好みにあった提示法にする必要がある. 転換点については時間情報のみを表にして提示する方法や転換点毎に区切った映像をタイムライン上に並べて示すといった方法がある. 情報の提示法には編集者の好みも影響してくるので, それに合わせて提示法を変えられるようにし, 編集者の好みに対応する必要がある.

図 3. システム全体の流れ



## 5. おわりに

今回は, 映像編集支援に役立つ話題転換点の抽出に向け, 特にインタビュー音声に着目し特徴分析し, ポーズ検出による候補点の検出をおこなった. しかしまだ十分な検出精度が得られないため,さらにポーズ検出の性能を高め, また, 別のアプローチからの転換点候補点抽出を行う必要がある. また, 候補点からの話題点抽出手法についても話者交替のみならず, キーとなる接続詞や相槌を抽出し, その位置関係から候補点を絞っていく等の方法の検討も必要である.

最終的には話題転換部, 不要箇所, 重要箇所情報を利用したシステムの構築を考える. そのためにはまず,重要部分の定義, 特徴分析が必要になる. 重要箇所に関しては発話の多さ(話の盛り上がり), キーワード出現率などから編集者の意図に合った重要部分の抽出を考えている. またそれらの情報をどう提示するかが編集の効率化に非常に重要になるため詳細な考察が必要である.

### 参考文献

[1]北出祐　南條浩輝　河原達也:談話標識と話題語を用いた重要文抽出手法の CSJ の学会講演における評価　話し言葉の科学と工学ワークショップ(2004)pp.61－66