Vol.2022-HCI-198 No.7
Vol.2022-EC-64 No.7
2022/6/16

動画視聴時の私的発話を可視化するシステムの提案

友広　純々野[1,a)]　塚田　浩二[1]

概要：近年，リモートワークやリモート学習など，一人で PC 作業を行う機会が増えている．例えば，オンライン会議はもちろん，少人数のオンライン対話や，講義動画の視聴なども盛んに行われている．本研究では，こうした PC 作業中の発話に注目し，PC 上で再生される音声の中でも動画上の音声と，その状況下で発せられる作業者の発話を同時に記録する．さらに，それらの発話を音声認識ライブラリを用いて文字起こしをし，準リアルタイムに発話内容を可視化するシステムを提案する．本システムを用いることで，動画視聴中の作業者の注目箇所を記録して振り返ったり，対話中の発話量をリアルタイムにフィードバックすることで発話機会を調整できる可能性がある．本稿では，システムの提案や実装について述べ，システムを利用した実験の結果，展望について述べる．

Visualize System for Private Utterance in Watching Video

Tomohiro　Suzuno[1,a)]　Tsukada Koji[1]

1. はじめに

他者との意思疎通を目的としない発話，いわゆる独り言を人々は日常の中で発している．本研究ではこの発話を私的発話と呼ぶ．

近年，意識的な私的発話であるセルフトークやライフログが注目されている．セルフトークとは自分自身と会話するという意味を持ち，感情をコントロールするメンタルトレーニングの 1 つ [1] とされている．例えば今川ら [2] は，セルフトークが脳活動を活性化させ，自信の向上や不安の解消といった精神に影響を与える働きがあると示唆している．また，外国語学習におけるシャドーイングが，学習者の聞き取り能力向上に有効な指導法であることが分かっている [3]．このような意識的な私的発話に加え，無意識的な私的発話も存在する．例えば，話者のひらめきや何か用事を思いだした時，話者自身の考えや感情を知らぬ間に発話しているものがある．

このように意識的な私的発話は，セルフトークやシャドーイングといった研究事例が存在するが，無意識的な私的発話を利用した事例は少ない．さらに私的発話は，発話者の記憶に残りにくいといった問題がある．

本研究では，動画を視聴 PC 等で動画を視聴するタスクを想定して，動画上の音声と，作業者の発話を同時に可視化するシステムを提案する．このシステムをユーザへ提示することで，私的発話への意識を高めることができ，発話の促しにつながるのではないかと考えた．

2. 関連研究

本章では，本研究に関連する研究事例として，「発話を用いた音声ライフログの事例」，「音声認識を用いた事例」の 2 つの観点から説明する．

2.1 発話を用いた音声ライフログの事例

Krosnick らは [4] は，PC で作成する成果物の制作プロセスを，発話情報を用いて作り上げる Think-Aloud Computing を開発した．このシステムを利用する際は，プロトコル法を応用し，成果物作成中の思考を発話しながら作業を進めるよう教示している．これにより作業を進めながら，制作プロセスを記録することができる．

川島ら [5] は，ユーザの発話を記録し，発話データを形態素解析，係り受け解析，概念解析することで，ユーザの

[1] 公立はこだて未来大学
Future University Hakodate
[a)] g2121040@fun.ac.jp

経験に関する情報を抽出している．この経験に関する情報を利用し，雑談内容を提示するシステムを構築している．このシステムを使って生成された雑談の提示は，ユーザに親近感を与える事を明らかにした．綾部ら [6] は，話者の発話時のピッチなどから感情を推測し，その感情タグを用いて過去に起こった出来事の分類を試みる手法を提案している．また，横山ら [7] は，話者が自覚しづらい口癖や言葉遣いを改善するための音声返戻システムを提案している．このシステムは利用者自身が発言を控えたい単語を設定し，発言中にその単語を発するとリアルタイムで発話音声が利用者にフィードバックされるシステムである．その結果，設定した単語や口癖に対する意識づけに効果があることを明らかにした．

私的発話を利用した研究事例として，長利ら [8] は，コーディング中の発話音声と文章を紐づけて，コードを読み返す際にその発話音声をフィードバックするシステムを制作している．コーディング中の発話音声と文章を紐づけて，コードを読み返す際にその発話音声をフィードバックするシステムを制作している．文字としては記録されなかった話者のアイデアや考察などを確認することを目指している．また，友広ら [9] は，日常生活における私的発話の傾向の予備調査を行い私的発話を分類し，私的発話の中でも，発話時の状況や行動に依存する私的発話と依存しないものに分類することができた．中でも，発話時の状況や行動を対する説明の頻度が高いと報告している..

発話を対象とした研究 [4]，[5]，[6]，[7] は多く取り組まれているが，私的発話を対象とした研究事例は少ない．また，友広ら [9] の研究は，私的発話の対象として，日常生活におけるさまざまな状況下で発せられるものを対象としていた．本研究では，日常生活の中でも動画視聴中という状況に限定したシステムを提案する．詳細については，3 章で後述する..

2.2 音声認識を用いた事例

現在，音声認識の精度は向上し，広く普及している．例えば，スマートフォンの文字入力や検索等でも，音声入力が標準で使用できるようになっている．さらに，さまざまな場面で音声認識を用いた文字起こしの事例が存在する．例えば，リアルタイムにテレビや講義の字幕を作成したり，議事録作成ツールなどに利用されている [10]，[11]．また，聴覚障害者のコミュニケーションを支援するために，他者の会話内容を音声認識で取得するといったアプリケーションの開発にも利用されている [12]．

3. 提案

本研究では，PC 等で動画を視聴するタスクを想定して，動画上の音声と，作業者の発話を同時に可視化するシステムを提案する．

PC 等で動画を視聴するタスクを選択した理由として，ユーザがある程度私的発話を行うと想定される点，視聴環境が安定していて音声記録が容易である点，動画自体の音声とユーザの発話を個別に記録して応用例等に活用できる点等を考慮した．なお，本論文では，動画自体の音声を「動画上の音声」，ユーザの発話を「作業者の発話」と呼ぶ．

次に，提案システムの概要について説明する（図 1）．まず，動画上の音声と作業者の音声を個別に取得し，音声認識サービスに入力する．次に，サービスから取得したそれぞれの音声の認識結果を用いて，発話量を示すグラフ表示と，発話内容を示す字幕表示の 2 種類の方法で視覚化を行う．

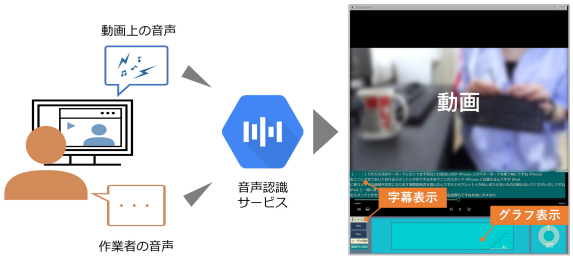

図 1 システムの概要

4. 実装

システム構成を図 2 に示す．実装には Python を用いた．また音声認識には，Google Cloud Platform の Speech-to-Text を用いた．Speech-to-Text を選択した理由は，リアルタイムに視覚化を行うために，音声ストリームを直接サービスへの入力として扱うことができるためである．

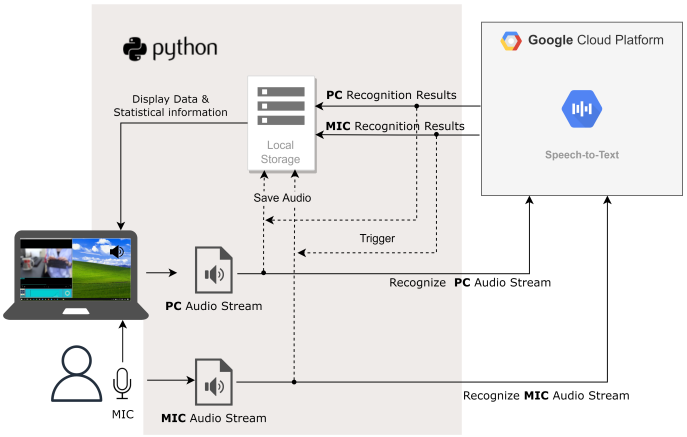

図 2 システム構成図

4.1 動画上の音声と作業者の音声認識と発話量の算出

本システムは，音声ストリームを Speech-to-Text へ入力し，そこから取得した認識結果を作業者へ可視化する．2 つの音声ストリームを分離して取得するため，プログラム

上で，動画上の音声は” ステレオ ミキサー (Realtek High Definit” を指定し，作業者の発話は” マイク配列 (Realtek High Definition” を指定した．次に，これらの音声ストリームに対して，以下の処理をそれぞれ適用して発話内容と発話量を取得した．

(1) 音声ストリームを Speech-to-Text に送信する

(2) Speech-to-Text から認識文字列 (発話内容) を取得する

(3) 認識文字列からをひらがなに pykakasi を用いて変換してカウントすることで，発話量を算出する．

4.2 可視化

可視化手法としては，発話量のグラフ表示と，発話内容の字幕表示の 2 種類を用意した．

4.2.1 発話量のグラフ表示

グラフ表示のスクリーンショットを図 3 に示す．画面左下のグラフ表示は Python の GUI ライブラリである PySimpleGUI を用いて実装した．これは大きく時系列に沿った発話量を示す折れ線グラフ（左）と，システムを起動してからの総発話量を示す円グラフ（右）を表示する．折れ線グラフの情報は，Speech-to-Text から認識文字列を取得する度に発話量を計算して更新され，左に行くほど古い時系列順に表示する．円グラフの情報は，PC 側／作業者側の発話量の割合を表示する．対象の期間は調整可能であるが，現時点では，システム起動後のすべての発話量の累積値を対象としている．

次に，各グラフの設計意図について説明する．折れ線グラフは，動画上の発話と作業者の発話のタイミングを俯瞰的に眺めることができるため，私的発話に基づくコンテンツの振り返り等に役立つのではないかと考えた．円グラフは，対象とする区間を適切に設定することで，定期的な私的発話を促すことができるのではないかと考えた．

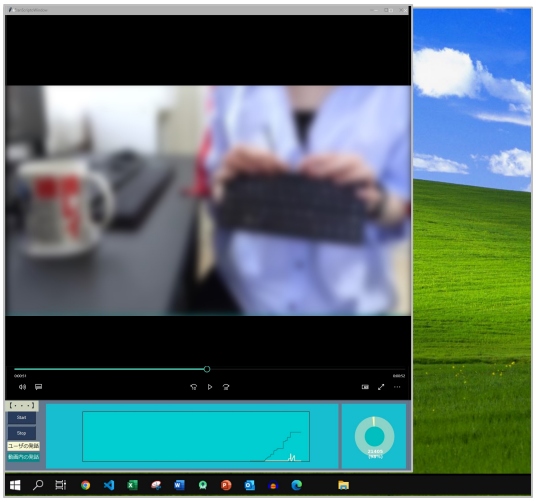

図 3 グラフ表示のスクリーンショット

4.2.2 発話内容の字幕表示

字幕表示のスクリーンショットを図 4 に示す．画面左側の字幕表示は Python の GUI ライブラリである Tkinter を用いて実装した．折れ線グラフと同様に，Speech-to-Text から認識文字列を取得する度に字幕を更新する．なお，グラフ表示と同様に，動画上の音声はネイビー，作業者の発話は白で表現されている．

字幕表示の設計意図としては，特に作業者の私的発話をリアルタイムに可視化することで，私的発話を記憶に残しやすくしたり，私的発話への意識を高めて発話しやすい環境を作る可能性があると考えた．

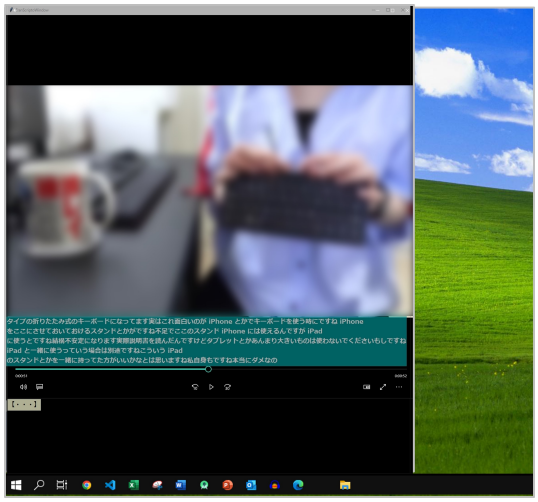

図 4 字幕表示スクリーンショット

5. 実験

ここでは，提案システムを動画視聴中に利用してもらい，視覚化手法の妥当性やシステムの効果について調査した．

5.1 手法

実験の被験者は，日常的に私的発話を行った経験があるものを対象とした．具体的には，事前に「普段，独り言を発言していますか」というアンケートに対して，3〜5 と回答した 6 名（大学生・男性 6 名) を選出した．実験の様子を図 5 に示す．実験は個室で行い，実験の前に「周りに人が誰もいないことを想定し，普段家で過ごしているときのように動画視聴してください」と教示した．被験者にはイヤホンとマイクを装着させ，27 インチのモニター上で動画とシステムを提示した．

実験は，システムでの情報提示手法が異なる 4 条件で行った（図 6)．条件 1 は，システム無しで動画のみを視聴させた．条件 2 は，グラフを提示しながら動画を視聴させた．条件 3 は，字幕を提示しながら動画を視聴させた．条件 4 は，グラフと字幕を動画を視聴させた．なお，条件 2,3,4 では提示内容が動画と重ならないように配置した．

条件ごとに 1 本 3 分 30 秒程度の動画を計 4 種類準備した．動画の内容は「おすすめのガジェットを数個紹介する」といった内容であり，動画内の人物は，常に発話を行っているものであった．また，動画の内容は情報系の大学生である被験者らにとって，一定の関心があると考えて選定した．なお，被験者には実験データの研究利用に対する同意を得た．

図 5 実験の様子

図 6 実験でのシステムの提示内容．左上から，条件 1〜4（条件 1: 提示なし．条件 2: グラフ．条件 3: 字幕．条件 4: グラフと字幕.）

次に，実験手順について説明する．実験は，以下の手順を各条件ごとに繰り返した．

（1） 動画とシステムの説明

（2） 動画視聴

（3） アンケートの回答

(1) では簡単に動画の概要を説明し，動画を止めずに最後まで視聴するように教示した．また条件 2〜4 の場合は，各条件で提示されるグラフや字幕について説明した．(2) で被験者が動画視聴中，実験者は被験者の目が届かないよう個室から退出した．実験中の様子は Open Broadcaster Software を用いて画面を撮影し，動画の音声と被験者の発話を録音した．

5.2 結果と考察

5.2.1 被験者の発話行動について

被験者 6 名のうち，動画視聴中に発話を行っている者は 2 名であった．なお，この 2 名は事前にアンケートで「普段，独り言を発話していますか」という 5 段階評価 (1: 発話しない〜5: 発話する) の回答に 4 または 5 と回答していた．発話を行った被験者を s1，s2 とし，各条件における発話量を表 1 に示す．どちらの被験者も発話内容は，s1，s2 共に動画に対する感想を述べていたが，動画の発話量（1600 以上）に対して，33〜54 程度と非常に少ない発話量であった．また，各条件で発話量に大きな差はなかった

ここで，実験前に被験者らの私的発話場面を回答してもらった結果を表 2 に示す．

表 1 各条件における発話量 (文字)

条件	2	3	4
システムの提示内容	グラフ	字幕	グラフと字幕
動画内の発話量	1646	1612	1846
s1 の発話量	54	48	54
s2 の発話量	33	39	34

表 2 設問：どのような場面で私的発話をしていますか．思い出せる範囲でご記入ください（自由記述）
※一部回答省略

被験者	回答内容
s1	ゲームをしながら考えているとき
s2	スポーツでミスをしたとき プログラミングをしているとき 食事の感想をいうとき
s3	コードを書くときや，なにか調べているとき
s4	プログラミングをしてるとき
s5	考え事をしているとき
s6	ゲームで緊張しているとき

発話を行った s1，s2 の発話量が少ないことに加えて，s3〜6 が全く発話を行わなかった原因として，今回の動画視聴タスクが，被験者が普段私的発話を行う場面と一致し

なかったことが一因と考える．本研究では，PC 上のコンテンツとしてガジェット紹介動画を選択したが，しかし，表 2 より被験者の私的発話場面は，ゲーム中やコーディング中，スポーツ中等の場面が多かった．今後は，こうしたコメントをもとに，スポーツやゲーム実況等の多様な動画や，コーディング等の多様なタスクを想定した検証を進めていきたい．

5.2.2 システムの利用方法

次に，提示した発話情報の利用方法として，システムの役立ち度合いを 5 段階評価で回答を得た結果を表 3 に示す．

表 3 設問: 視覚化された発話情報は，役立つと感じましたか．5 段階評価，1: 感じなかった 〜 5: 感じた

条件	2	3	4	
提示内容	グラフ	字幕	グラフ	字幕
平均 (標準偏差)	3.2(1.0)	3.3(1.5)	2.8(1.2)	3.3(1.5)

表 3 より，条件 2 で提示されたグラフの役立ち度の平均は 3.2 となり，肯定的，否定的な意見の双方が挙げられていた．肯定的な意見として「リアクションの起きる場所の特定に役立つと感じた」といったものがある一方，「用途が分かりにくかった」といった意見も挙げられていた．

次に，条件 3 で提示されたグラフの役立ち度の平均は 3.3 となり，こちらも肯定的，否定的な意見の双方が挙げられていた．肯定的な意見として「動画内で紹介された固有名詞や聞き逃した箇所の確認ができた」がある一方，否定的な意見として「動画の音声も聞きやすかったので，あまりメリットを感じなかった」といった意見が挙げられていた．被験者らは，本システムを動画に表示される字幕機能として認識していた可能性がある．

条件 4 の（グラフ/字幕）による提示は平均（2.8/3.3）とグラフ表示の方が役立ち度がやや評価が低かった．この理由としては，2 つの情報を同時に可視化すると，情報過多となるためだと考えられる．実際に被験者から「画面のサイズが大きかったため，両方（グラフと字幕）を見ることが難しかった」,「字幕の方のみを追っていた」などの意見が挙げられていた．

5.2.3 システムの煩わしさ

システムの提示によりコンテンツの視聴を阻害することがないか調査するため，システムの煩わしさを 5 段階評価で回答を得た（表 4）．

表 4 設問: 視覚化された発話情報は，煩わしくないと思いましたか 5 段階評価，1: 思わなかった 〜 5: 思った

条件	2	3	4	
提示内容	グラフ	字幕	グラフ	字幕
平均 (標準偏差)	3.8(1.2)	3.8(1.2)	3.5(1.8)	4.2(1.8)

まず，条件 2 で提示されたグラフの煩わしさの平均は 3.8，条件 3 で提示された字幕への煩わしさの平均は 3.8 であった．続いて，条件 4 の（グラフ/字幕）による提示は平均（3.5/4.2）とシステムの煩わしさは字幕表示の方が若干煩わしさを感じないという結果になった．2 つの情報を同時に提示しても，コンテンツを邪魔する可能性は低いが，グラフ表示に関して「ほとんど見ていなかった」といった意見が見られた．

6. まとめ

本稿では私的発話の利用方法を探る目的で，作業者の発話と動画上の音声を可視化するシステムを開発した．このシステムを私的発話を行う被験者に対して，動画視聴時に提示し，5 段階の主観評価によりシステムの役立ち度/煩わしさに関する回答を得た．結果，私的発話を発する傾向を持つ被験者を対象としたものの，発話者は 6 名中 2 名であり，作業者の発話を視覚化したときのデータは多く集まることはなかった．しかし，提案手法によるメリットやデメリットが意見として挙げられたことに加えて，私的発話の発生場面を再検討する必要があることが分かった．今後は動画視聴以外の場面として，私的発話が発生するような場面を再検討して実験を行い，私的発話を活用できるシステム開発につなげていきたい．

参考文献

[1] 高妻容一：基礎から学ぶ!メンタルトレーニング，ベースボール・マガジン社，2008.
[2] 今川 新悟，松本 清，佐久間春夫：セルフトークの精神生理学的効果について，バイオフィードバック研究，第 45 巻，第 1 号，pp. 25-32，2018.
[3] 玉井 健：シャドーイングの効果と聴解プロセスにおける位置づけ，時事英語学研究/日本時事英語学会 [編]，第 1997 巻，第 36 号，pp. 105-116，1997
[4] Rebecca Krosnick, Fraser Anderson, Justin Matejka, Steve Oney, Walter S. Lasecki, Tovi Grossman, George Fitzmaurice：Think-Aloud Computing: Supporting Rich and Low-Effort Knowledge Capture. CHI '21, 199, pp.1–13, 2021.
[5] 川島 嵩弘，西村 隆志，安江 駿亮，六沼 元貴，和田 史織，杉本 徹：ライフログ雑談対話システムに関する研究，人工知能学会研究会資料，pp.120-121，2016.
[6] 綾部 櫻子，田野 俊一，市野順子，岩田 満，橋山 智訓：イベントの内容，感情をロギングするリッチなサウンドライフログの提案，第 28 回ファジィシステムシンポジウム，pp. 625-630，2012.
[7] 横山 紗菜，鈴木優：言葉遣いを改善する音声返戻システムの開発，情報処理学会，インタラクション 2020，pp.936-939，2020.
[8] 長利 槙吾，寺田実：独り言と文字を紐付けて自動記録するテキストエディタ，情報処理学会，インタラクション 2016，pp.710-712，2016.
[9] 友広 純々野, 塚田 浩二：日常生活における私的発話識別法の提案 ，情報処理学会研究報告，第 194 回 研究報告ヒューマンコンピュータインタラクション (HCI), pp.1-6, 2021.
[10] 今井亨，奥 貴裕，小林彰夫：音声認識によるリアルタイム字幕放送の進展，情報処理学会研究報告 第 88 回，

pp.1-pp.6，2011
[11] 河原 達也，秋田 祐： 哉聴覚障害者のための講演・講義の音声認識による字幕付与，日本音響学会誌 74 巻 3 号，pp.156–pp.162，2018.
[12] 服部 哲，柴田 邦臣：聴覚障害児・者のコミュニケーションを支援する Android アプリの開発，ワークショップ 2014（GN Workshop 2014）論文集，pp.1–pp.6，2014.