動画への自動テロップ付けシステムと応用

張晋瑜[†1]　神場知成[†2]

概要：従来の字幕やキャプションは外国語の翻訳や，聴覚の障がい等で音声が聞き取りにくい人に対して映像の内容を理解しやすくために利用されてきた．一方，日本のバラエティ番組などにおいて，視聴体験を拡張する目的で画面上の見やすい位置にデザイン化した文字等を表示するテクニックはテロップと呼ばれ，英語ではインパクトキャプションと呼ばれるようになってきている．従来，字幕の自動生成等の研究はあるが，それらとは異なり本稿では，テロップの自動生成システムに関する初期的な試みを示す．映像から音声認識を利用してテロップを自動表示し，66 名の大学生を対象としてアンケート評価を行った結果，テキストおよびアイコンの両方を用いたテロップが「おもしろい」「わかりやすい」「見やすい」いずれの項目でも高い評価を得た．自動テロップ付けは，音声認識だけでなく感情の自動分析などを利用しても良く，映像の新しい楽しみ方を提供できる可能性がある．さらに，自動テロップ付けをオンラインビデオ会議システムに適用することを提案し，プロトタイプを示した．

1. はじめに

字幕は，日本では外国語を日本語に翻訳表示するために使われる場合が多いが，米国等における字幕放送（クローズドキャプション）は，聴覚障がいや言語能力的な理由で音声を聞き取りにくい人に対してニュース，ドラマなどのナレーションやセリフを映像と同じ画面に表示するサービスである．音声情報を視覚情報に変換し，映像を補足することで視聴者が内容を理解しやすくなり，視聴エクスペリエンスを向上することにつながる．また，字幕以外の環境音や発話者の名前などの説明文を含むこともあり，映像の内容を理解するための補助手段として利用されている．

一方，日本では翻訳用の字幕とは別に，テレビ番組にテロップが広く用いられている．ここでいうテロップはテレビ映像に重ねる形でテキスト等を表示するものを指し，視聴者にインパクトを与えることやエンタテインメント性を主目的として，デザインやレイアウトも工夫している．このような表現方法はもともと 1990 年代に日本のバラエティ番組から始まったと言われており，現在では特にアジア地域で広く用いられ，欧米等にも広がり始めている[1][2].

なお，日本ではテロップと呼ばれることが多いが，上記に述べたような聴覚障がい者補助や翻訳を目的としたものと区別するため，海外ではインパクトキャプションと呼ばれるようになっている．

本稿は，インパクトキャプションの自動生成の試みである．従来から，一般に字幕作成には手間がかかるため，作成の自動化や補助を目指す研究は行われている[3-9]．これに対して本稿は，インパクトキャプションを目的とするため，手法や効果が大きく異なる．本稿では複数の表示方法の比較，その応用可能性などについて述べる．

2. 従来の研究

主に聴覚障がい者のための字幕作成を自動化あるいは補助する技術の研究はさまざまなものがある．Hong ら[4]は，音声情報を可視化する際に映像の発話者を認識しやすくように，適切な位置にスクリプトを配置することで，映像アクセシビリティの向上を実証した．Brown ら[5]は聴覚障がい者を対象としてダイナミック字幕のテストをしてアイトラッキングデータを収集し，ダイナミック字幕がユーザーエクスペリエンスの向上につながることを実証した．そこでは従来の字幕と比べ，ダイナミック字幕を利用したユーザーの視線は字幕なしの時に近いという結果が得られている．Hu ら[6]は，画面上の字幕を発話者の横に配置する表示方法を提案し，ユーザビリティ調査の結果，視聴エクスペリエンスの向上と視聴疲労の軽減を示した．江草ら[7]は，吹き出し字幕の有効性を明らかにするため，健聴者に対して人形劇の鑑賞に字幕付き映像を用いて評価実験を行い，吹き出し型字幕は視覚的負荷の軽減効果が期待できるとしている．さらに，字幕の表示量やサイズ，表示位置の制御などの課題があるものの，演劇のような脚本がない講義や，討論場面に吹き出し型字幕提示法を拡張できると述べている．また Jiang 等[8]は，聴覚障がい者を対象としてキャプションを表示する際に，利用者が画面上のどこを見ているかを視線分析によって検知し，その妨げにならないように表示するシステムを示している．Amin 等は[9]，聴覚障がい者を対象としてキャプションを表示する際に，キャプション表示位置を 4 種類（画面の上方，上 3 分の 1，下 3 分の 1，画面の下方）に設定して，それによる後方画面の隠れ（オクルージョン）が，視聴者が受け取る全体的な印象にどのように影響を与えるかをニュース番組，インタビュ

†1 東洋大学 情報連携学研究科
†2 東洋大学 情報連携学部

ーやトークショーなど 6 つのジャンルに分けて検討し，定式化している．そこではオクルージョンが隠す対象や量・時間がキャプション全体としての印象に影響を与えることを示し，たとえばニュースであれば，話し手の目，ニュースにおける現在のトピックなどが隠れることの影響が大きいことが示されている．

インパクトキャプションの自動生成に関する研究は見当たらないが，インパクトキャプションの効果や役割に関する議論は Sasamoto が行っており[1][2]，インパクトキャプションが日本のテレビ番組で 1990 年代以降に広がり，現時点では主にアジアに普及し，欧米にも広がり始めていることを指摘し，その発展の背景，歴史，目的等を事例とともに詳細に分析している．そこではインパクトキャプションが，ユーモア効果を生み出すためや，登場人物への共感と理解を深めるためなどさまざまな目的で意図的に利用されていると述べている．

3. 提案する自動テロップ付けシステム

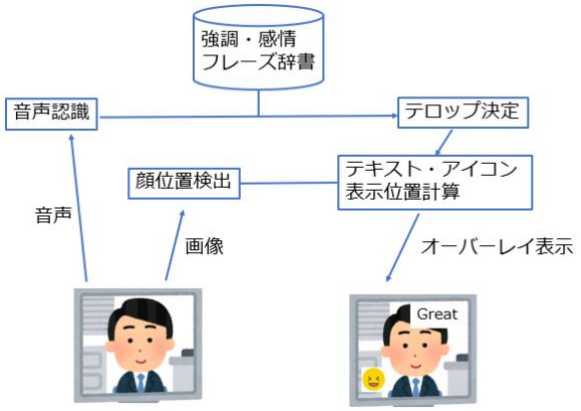


図 1. システム構成

図 1 に筆者等が作成したシステムの構成を示す．システムは Web 上に実装しており，全体の流れは次のようになる．

1) 入力映像からの顔検出と，音声認識
2) 音声認識結果にもとづき，テロップ表示を決定
3) 映像にテロップをオーバーレイして表示

1)において，入力映像のうちで画像から，OpenCV ライブラリを用いて顔の位置検出を行う．入力映像のうちで音声は Web Speech API の SpeechRecognition API を用いて認識し，テキストに変換される．入力映像としてあらかじめ録画した映像を使う場合，Web Speech API は直接パソコンのマイクから音声を拾うため，ノイズとハウリングの影響を防ぐために，Loopback を用いてサウンドルーティングを行った．なお同 API を用いた場合，音声認識結果が，発話の切れ目（一定の時間的な空白）までは未確定状態になり，空白時点で確定されるが，本システムでは表示のリアルタイム化のために未確定状態のテキストを利用している．

2)では，あらかじめ作成してある強調・感情フレーズの辞書を用いてテロップを決定する．現時点で強調・感情フレーズ辞書は限られた内容の会話におけるシステムの動作確認のために便宜的に作成したもので，たとえば「Yesterday」「sushi（寿司）」のような単語を強調フレーズ，「happy」「disaster」のような単語を強調フレーズと感情フレーズの両方に登録している．強調フレーズはテキストテロップ，感情フレーズはアイコンテロップとして表示することにしており，両方に登録されている場合は両方が表示される．

3)では，映像にテロップをオーバーレイ表示する．この際，テキストテロップは顔の左上をスタートとして配置し，アイコンテロップは顔のやや左横に配置している．これはテロップが目や口を隠さないように便宜的に決定したものである．なお，現時点でシステムは英語音声のみに対応しているが，この理由は，OpenCV の JavaScript ライブラリを用いた画面表示で日本語が文字化けしたためであり，本質的な理由はない．

システム動作中の画面例を図 2 に示す．これはいずれも，テキストとアイコンの両方のテロップが表示されている．

(a) 画面例 1（excited という単語に反応）

(b) 画面例 2（disaster という単語に反応）

図 2. システムの動作例

4. 評価と考察

4.1 評価方法

オリジナル　　movie1: テキストのみ

movie2：アイコンのみ　　movie3：テキストとアイコン

図 3. 比較した表示方法

システムの有効性を評価するために，66 名の大学生を対象とし，アンケート調査を行った．

評価対象としたのは，図 3 左上に示すオリジナル映像を基準としたときに，それに対してテキストテロップ，アイコンテロップ，両方のテロップをオーバーレイ表示した movie1～3 である．映像は 22 秒間で，次のような発話をしている．movie1～3 は，オリジナル映像をシステム上で再生してテロップをオーバーレイ表示した映像をあらかじめ録画して，その映像を再生することで実験を行った．

発話文：

Yesterday, I went to a **sushi** restaurant. I was very ***excited*** because **sushi** is a very ***famous*** Japanese food. I tried many kinds of **sushi** and the taste was ***great***. But I ***accidentally*** put too much **wasabi** and that was a ***disaster***.

ここで，強調フレーズに登録されている単語は太字，感情フレーズに登録されている単語はイタリック体で表示している．両方に登録されている単語は太字のイタリック体である．

評価は，上記の 66 名全員が同じ教室にいて，映像は教室正面にプロジェクターで大きく投影し，音声は部屋のスピーカーから十分な音量で流した．なお，学生のほとんどは日本人であり，数名，「日本語が母国語ではないが，通常の講義は常に日本語で受けている」という学生が含まれる．ただし，彼（女）等はすべて，英語は母国語ではなく，英語力は日本人学生とほぼ同等である．

具体的な評価手順は以下の通りである．

1) 最初にオリジナル映像を 2 回見せる．
2) 次に，オンラインでアンケート画面を提示し，「これから 3 つの映像を見せるので，オリジナル画像と比較して回答してほしい」と伝える．アンケート画面では，テロップ表示をする 3 つの映像それぞれに対して，図 4 に示すように「おもしろさ」「わかりやすさ」「映像の見やすさ」を評価するようになっており，オリジナル映像を 0 としたときに，+1 または+2（よりおもしろい，わかりやすい，見やすい），-1 または-2（その逆）などから選択できるようになっている．3 つの映像に対する評価欄は 1 つのページに表示されており，別の映像を見たあとで，既に見た映像の評価を修正することも可能である．また，3 つの映像に対する回答欄の最後に，「すべての映像を見た上での感想」という自由記述欄を設けている．
3) 上記の説明をしたうえで，movie1（テキストテロップのみ），movie2（アイコンテロップ），movie3（両方のテロップ）それぞれの映像を，順番に 1 つあたり 2 回見せる．回答者は映像を見ながら適宜，アンケートに記入し，記入し終わったら提出する．

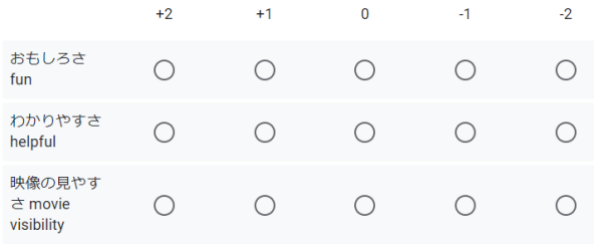


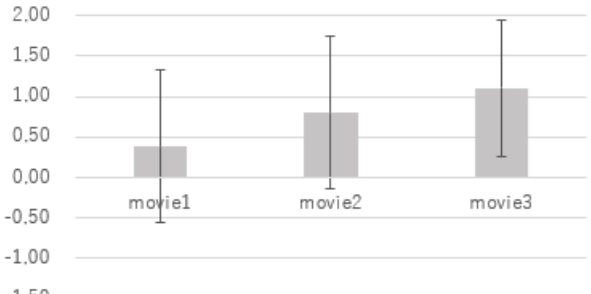
図 4. アンケート画面

4.2 評価結果

評価結果を図 5 に示す．概要は下記のとおりである．

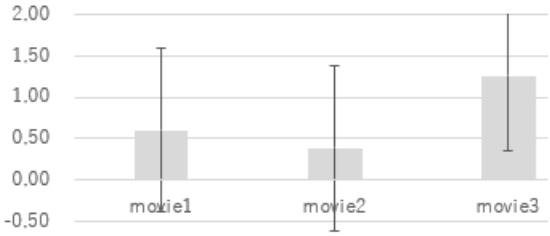
- いずれの評価においても，movie3（テキストとアイコンの両方を表示）が一番高い評価となっている．
- movie1 と movie2 を比較すると，わかりやすさでは movie1 の方が評価が高く，おもしろさでは movie2 が高くなっている．つまり，テキストはわかりやすさ，アイコンはおもしろさに貢献している．
- 見やすさの観点では，全体的に平均の数値が低い．映像間の比較では movie1 < movie 2 < movie3 となっているが，特に movie1 では，見やすさの平均がマイナス値（-0.28）となっている．つまり，テキストが重ね表示されることでオリジナル映像よりも見づらいという評価である．movie2 は平均でプラスになっているが，movie2 においてもアイコンによって映像の一部は隠れてしまっているので，「見やすくなった」というよりも，「アイコンが表示されることの効果により，画面が隠れることの欠点が補われて結果的にプラスになった」と考える方が妥当であろう．

次に，アンケートの自由回答欄の主なコメントは次の通りである（コメントは一部編集）．



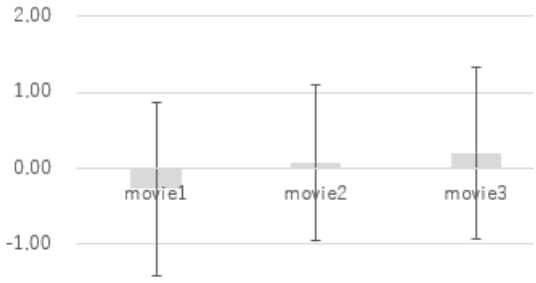
	1	2	3
AVG	0.38	0.80	1.11
SD	0.94	0.95	0.84

(a) おもしろさ



	1	2	3
AVG	0.60	0.38	1.26
SD	0.99	1.00	0.92

(b) わかりやすさ



	1	2	3
AVG	-0.28	0.08	0.20
SD	1.14	1.03	1.13

(c) 見やすさ

図 5. 評価結果

1) 肯定的なもの
・テロップだけで非常にわかりやすくなることに驚いた
・話すタイミングに合わせてテロップ表示されるのが良い
・感情を表すアイコンのおかげで英語が理解しやすい
2) 否定的なもの
・簡単な感情はアイコンを表示しなくてもわかる
・アイコンが邪魔になる，位置が悪く大きすぎて見にくい
・表示のタイムラグがある
・テロップの位置がちらちら動くのが気になる
3) その他
・たとえば sushi（寿司）なども，アイコンにすると良い
・フォント，色などをもっと工夫するとおもしろい
・テキストだけは無機質で，アイコンだけはわかりにくい

5. 考察

5.1 実用性

評価を全体的にみると，テロップは好意的に受け取られており，特にテキストとアイコンの両方でテロップを表示することには効果があったと考えられる．通常の動画をこのようなシステムで用いることで，同じ動画をより楽しむことができる可能性がある．

ただし，今回のテキスト，アイコン表示はいずれも事前に発話内容を決めて，それにもとづいて何をテロップ表示するかを決めたものである．たとえば restaurant はテキスト表示していないのに対し wasabi はテキスト表示をするなど一貫性があるとは言えず，あくまで表示方式の妥当性を簡易調査した段階にとどまっている．これらについては今後，固有名詞辞書，感情辞書の利用が考えられる．また，必ずしも認識結果をそのままテキスト表示する必要はなく，言い換え辞書などを作成しても良い．さらに，本稿では音声認識だけを用いてテロップ表示の判断を行ったが，参加者の声のトーン，顔の表情などから推定した内容にもとづいてテロップ生成を行うことも考えられる．

また，テロップ表示場所，テロップ表示タイミングについても検討すべき点が多々残されている．画面内に複数の発話者がいる場合に話者を認識して表示を仕分けるなどの方法もできていない．既存研究にもあるように，表示場所，表示されている時間は画面の見やすさに大きな影響を与える．本システムでは，テロップ表示すべき次の単語が出現した時点で前のテロップが消えて上書きしており，きわめて短時間でテロップが消える場合もある．さらに，前述のようにテロップ表示位置は顔の位置をリアルタイム検出した点からの相対位置で決定しているので，顔の移動に伴いテロップも頻繁に動くという状態となっており，自由回答にもあった「ちらちら動く」という感想につながっていると考える．これらはいずれも今後の検討課題である．

5.2 オンライン会議システムへの応用

図 6. オンライン会議システムへの応用例

最後に，本システムの応用可能性として，オンラインのビデオ会議を挙げる．コロナ禍でオンラインによるビデオ会議は急増しており，オンライン会議をより便利に利用するための様々な検討が行われている[10]．しかし，利便性だけではなく，対面会議とちがってオンライン会議では緊張して話しづらくなる等の状況も発生している．たとえば，テキストメッセージにおける絵文字はコミュニケーションを大幅に柔らかいものにし，テキストだけのやりとりによる人と人とのコミュニケーションを円滑することに役立っている．ビデオ会議においても，本システムのようなテロップを利用することにより，人と人とのコミュニケーションを柔らかいものにできる可能性があると考え，利用を検討中である．図 6 は，オンラインミーティングの場面で利用した場合のプロトタイプである．これについても，全員が発話する 20 秒程度のオンライン対話映像を作成し，そこにテロップ表示を重ねる実験を行ったみたところ，単なるビデオ映像だけを表示しているよりも見ていて好感の持てるものになるという印象を得た．

オンライン会議システムについては従来，筆者等は参加者の顔画像から参加者の感情推定を行うシステムの構築も行っており[11]，そのような感情推定と組み合わせることで，より柔軟なテロップやアイコン生成が可能になると考えられる．

6. おわりに

動画への自動テロップ付きシステムを提案し，試作に対するアンケート調査により，システムの基本的な有効性を確認するとともに，通常の動画視聴だけでなく，オンライン会議への応用も示した．今後テロップの表示方法について，さらに検討を行う予定である．

謝辞 本研究は，東洋大学井上円了記念研究助成および東洋大学重点研究推進プログラムにより助成を受けたものです．同助成に感謝いたします．また，実験およびアンケート評価に協力してくれた東洋大学情報連携学部各位に感謝いたします．

参考文献

[1] R. Sasamoto: Impact caption as a highlighting device: Attempts at viewer manipulation on TV, Discourse, Context and Media Vol. 6, pp.1-10 (2014)
[2] M. O'Hagan and R. Sasamoto: Crazy Japanese subtitles? Shedding light on the impact of impact captions with a focus on research methodology, Eyetracking and Applied Linguistics (pp.31-58), Chapter 3 (2016).
[3] 河原達也,秋田祐哉，聴覚障がい者のための講演・講義の 音声認識による字幕付与，日本音響学会誌, Vol.74,No. 3, pp. 156-162 (2018)
[4] R. Hong, et al. Dynamic captioning: Video accessibility enhancement for hearing impairment, MM '10, pp.421-430 (2010)
[5] A. Brown, et al. Dynamic subtitles: The user experience, TVX 2015, pp.103-112 (2015)
[6] Y. Hu, et al. Speaker-following video subtitles, ACM Transactions on Multimedia Computing, Communications and Applications, pp.1-17 (2014)
[7] 江草遼平 ほか，視線計測装置を用いた吹き出し型字幕提示法の視線移動量低減効果に関する有効性の評価，ヒューマンインタフェース学会論文誌, Vol.21, No.4, pp.381-390 (2019)
[8] Bo Jiang, Sijiang Liu, Liping He,WeiminWu, Hongli Chen, and Yunfei Shen. 2017. Subtitle Positioning for E-Learning Videos Based on Rough Gaze Estimation and Saliency Detection. In SIGGRAPH Asia 2017 Posters. Article 15 (2017)
[9] A.A. Amin, S. Lee, M. Huenerfauth: Watch It, Don't Imagine It: Creating a Better Caption-Occlusion Metric by Collecting More Ecologically Valid Judgments from DHH Viewers, CHI '22, Article No. 459, pp. 1-14 (2022). https://dl.acm.org/doi/abs/10.1145/3491102.3517681
[10] S.Samrose, et al.: MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings, Proceedings of the 2021 CHI, Article No. 252, pp.1-13 (2021)
[11] 神場：オンライン会議システムにおける感情推定と参加者へのフィードバック方法，インタラクション 2022, 6D18 (2022)