映像視聴体験を向上するための テロップ自動付加システムの開発と評価

張晋瑜[†1]　神場知成[†2]

概要：さまざまな動画視聴サービスや映像を用いたオンライン会議の増加など，一般利用者が動画を視聴する機会が急増している．従来，聴覚障がい者や外国語視聴者などへのサポート機能として字幕表示があるが，それとは異なるものとして，エンタテインメント番組等で特に日本をはじめとするアジア地域で広がっているテロップ表示がある．本稿では，動画に対して自動的にテロップを重畳するシステムを提案する．動画内の音声データの認識結果をもとにテキスト，イメージなどを用いてテロップ重畳を行うシステムを開発して評価を行ったところ，テキスト表示はわかりやすさ，イメージ表示はおもしろさなどの向上に有効な可能性が示された．基礎的な評価実験にもとづき，辞書機能，テロップ表示位置，表示時間の調整などを行った．今後は，音声だけでなく映像の認識結果にもとづくテロップ表示や，既存の動画視聴だけでなくオンライン会議における対話画面への重畳などについても検討を進める予定である．

1. はじめに

世界中で動画コンテンツの視聴が急速に増加している．従来のテレビや映画館での視聴とは異なり，新型コロナウイルスで自宅にいる時間が増えることなどもあり，YouTube を代表とするオンライン動画プラットフォーム，Netflix や Amazon Prime などの有料動画ストリーミングサービスの利用者が急増している．インプレス総合研究所発行の「動画配信ビジネス調査報告書 2023」によると，2023 年に有料動画配信サービスの利用率は 31.7%となり，前年度調査から 2.8 ポイント増加している．また，視聴者の動画再生方法はますます多様化しており，同報告書で有料動画配信サービス利用者に視聴に利用するデバイスについて尋ねたところ（複数回答），テレビで視聴するユーザーが 53.6%となり，昨年調査の 48.8%から 4.8 ポイントと増加した．またパソコンは 46.9%，スマートフォンは 44.4%とどちらも高い比率を占めており，外出中の電車やバスなどでの移動中にも利用されていると考えられる．

さて，動画コンテンツの視聴を補助するテクニックの一つとして字幕表示がある．これは聴覚障がい者に対するサポート機能として用いられる場合や，外国語視聴者に対する翻訳表示の役割を持ち，デジタルデバイド緩和策の一つと言える．

一方，日本を起点として主にアジア地域では，エンタテインメント目的でテロップ表示が用いられている．これは，たとえば出演者のセリフを補完，強調して盛り上げることが目的であり，字幕と比較したときに画面上の表示位置，表示方法のバリエーションが大幅に大きい．

本稿では，映像視聴体験の向上を目指し，テロップ自動付加システムを提案し，その有効性の基本的な検証を行う．なお，類似した用語として字幕，クローズド・キャプション，テロップ等があるが，ここでは聴覚障がい者への対応や外国語視聴者への翻訳に対応するためものをキャプションまたは字幕と呼び，エンタテインメント目的のものをテロップと呼ぶこととする．なお後述するように，後者のものを，より正確にインパクトキャプションと呼んでいる事例もある．

2. 従来の研究

従来の字幕やテロップに関する研究は，聴覚障がい者向けの字幕作成の自動化や補助技術に焦点を当てたものが多いが，テロップに関するものも行われている．

2.1 表示位置やレイアウト

自動字幕配置技術に関するものとして，表示位置によってオクルージョン（重畳による後ろの物体の隠れ）による視聴者の印象が変化することが指摘されている．Amin 等は[1]，聴覚障害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，定式化した．そこではオクルージョンが隠す対象・量・時間によって視聴者への印象が変わることを示し，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場合に影響が大きいことが示されている．

Hong 等が提案した Dynamic Captioning は[2]，顔検出と認識，ビジュアル顕著性分析，テキストと音声のアラインメントなどの技術を用いて，映像のアクセシビリティを高

†1 東洋大学 情報連携学研究科
†2 東洋大学 情報連携学部

めようとしている．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．

2．2 視線分析

Brown 等は[3]聴覚障害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の移動は字幕なしの時に近いという結果が得られている．

Jiang 等は[4]，聴覚障害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．また，Hu 等は[5]，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．

2．3 吹き出し型字幕やテロップ

江草らは[6]，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，表示における人形への自動追従機能が吹き出しの内容を理解することや話者の特定に役立つことや，感情を表現する吹き出し機能が聴覚障害を持つ人々がキャラクターの感情を理解するのに助けになることなどを示している．これは「字幕」との位置づけだが，単なるテキスト表示ではなく視聴効果を狙った側面もあり，いわゆるテロップに類似する面もある．

テロップについては，Sasamoto が「インパクトキャプション」と呼んで議論を行っている[7][8]．インパクトキャプションは 1990 年代以降に日本のテレビ番組から広がり，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と指摘し，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，登場人物への共感や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．本稿は，このようなインパクトキャプションの自動生成を目指すものである．

3. テロップ自動作成システム

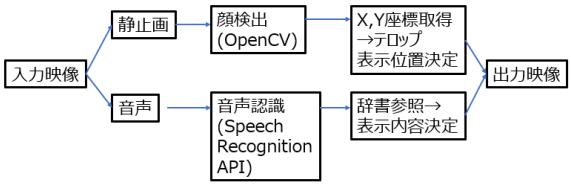


図 1．システム構成

3.1 提案システム

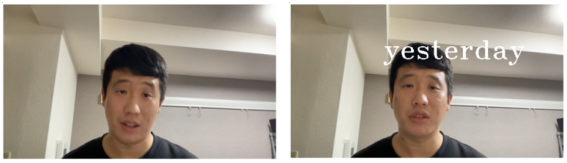
図 1 に基本的なシステムの構成を示す．システムは Web 上に実装しており，全体の流れは次のようになる．

1) 入力映像からの顔検出と，音声認識
2) 認識結果にもとづき，テロップ表示内容や位置を決定
3) 映像にテロップをオーバーレイ表示

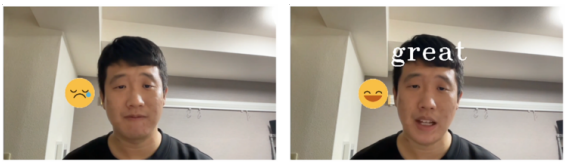
なお，これは現時点での実装にそった構成であり，将来的には音声認識だけでなく映像認識を用いる等，付加的な要素も検討中である．

3.2 基礎評価

テロップ付加というコンセプトの有効性を評価するために，簡易実装したシステムを用いて作成した映像で 66 名の大学生を対象とし，アンケート調査を行った．



オリジナル　　movie1: テキストのみ



movie2 : イメージのみ movie3 : テキストとイメージ

図 2．　比較した表示方法

評価対象としたのは，図 2 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，イメージテロップ，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 秒間で発話音声が含まれているがここでは省略する．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い，「わかりやすさ」「おもしろさ」（いずれも 1～5）等の評価および自由回答によるアンケートを行った．

3.3 評価結果

- いずれの評価においても，movie3（テキストとイメージの両方を表示）が一番高い評価となった．
- movie1（テキスト）と movie2（イメージ）を比較すると，わかりやすさでは movie1 の評価が高く，おもしろさでは movie2 が高くなった．つまり，テキストはわかりやすさ，イメージ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 となったが，特に movie1 では，見やすさの平均がオリジナル映像と比較してマイナス値（-0．28）となった．movie2

は平均でプラスになったが，movie2 においてもイメージ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，「イメージの効果が，画面が隠れることの欠点を補って結果的にプラスになった」と考える方が妥当であろう．

この他，自由回答として「テロップ表示（テキストおよびイメージ）が常に動いて見づらい」等の意見もあった．

3.4 辞書の設計

基礎実験により，テロップ表示においてテキスト，イメージそれぞれが異なる効果を持つことが確認できたので，実用性の向上を目的とした開発を行った．

まず，発話されたテキストに対応してテロップとして表示する可能性があるテキストやイメージを指定するための辞書設計を行った．そこではテキストのフォント，フォントサイズ等も設定可能としている．現在の辞書の一部を図 3 に示す．

単語	表示テキスト	表示画像	フォント	フォントサイズ
Japan	Japan	jpn-png	FONT_HERSHYE-SIMPLEX	1.3
amazing	amazing	amazing.png	FONT_HERSHEY_SIMPLEX	2
wow	wow	wow.png	FONT_HERSHEY_COMPLEX	2
...	...	...	...	..

図 3. 辞書の一部

3.5 テロップ表示位置の制御

基礎実験に対するアンケートでは「テキストとイメージの位置が常に変動している．テロップのテキストが読みにくい」というコメントがあった．これは，OpenCV によりフレームごとに顔の検出を行い，その横にテロップ表示をしているせいだと考えらえる．つまり，発話者が話している間は顔が動くと，取得される位置座標（X と Y）もフレームごとに変化している．この問題を解決するために，テロップの表示位置を一定時間は固定するようにした．

音声認識の結果から単語を抽出し，辞書にマッチする単語がある場合にテロップを表示するようにしているが，単語が一致してから 2 秒間は取得した位置座標を更新せず，その表示位置を一時的に固定する．

また，基礎実験ではテロップが顔の上に表示される場合があったが，テキスト表示位置が必ず顔の右横となるように修正をした．

3.6 テロップ表示時間の制御

さらに別の問題として，発話中にテロップ表示に対応する単語が続けて出現すると，前に表示されたテロップをすぐに上書きしてしまうため,「一瞬表示されただけで上書きされるテロップ」が発生していた．テロップの表示時間は視聴者がテキストを読み取るのに適切な時間を与えるよう調整されるべきであり，文字数や内容の適切なバランスを考慮しつつ，理解しやすく，追いつきやすい速度で字幕が表示されるよう設計する必要がある．

今回，音声認識の結果からテロップを表示する際に，表示される時間を調整する機能を追加した．これにより，テロップが適切な間隔だけ表示され，視聴者が画面を適切に読み取り,情報を取得しやすくなるよう配慮した(ただし，上書きされる問題は解決していない)．

今テレビ上や映画の字幕の読み上げ平均秒数は，通常，言語や文の長さ，字幕の表示スタイルなどによって異なっている．一般的には，読み上げられる文字数に基づいて平均的な表示時間が設定され，最も一般的なガイドラインとしては，1 つの字幕の表示時間が 2〜7 秒程度であり，また日本語は 1 秒間に 4 文字以内，英字（アルファベット）は 1 秒間に 12 文字以内というのが，一般的な字幕の速度範囲である．

今回は，改良後のテロップ表示時間を字幕の読み上げ平均秒数の最も短い 2 秒に設定した．ただしこの数値は一例であり，テロップの内容と実際の設定は特定のコンテンツや視聴者のニーズによって異なる場合があるため制御手法は今後の課題である．

3.7 システムによる画面表示例

ここでは，初期の評価用プロトタイプに対して上記のような強化を行ったシステムの動作例を示す．リアルタイムの表示性を重視して画像サイズは小さめの 540×360 に，コーデックは H.264 に設定した．映像の長さは 27 秒であり，オンライン会議のような場面で二人が対話する形であるが，現在の実装では音声からの話者認識などの処理を行っていないため，同時に二人が話す画面は利用していない．

話者は次のような発話を順番に行っている．

Yoko: Hi John, how was your trip to **Japan**?
John: Oh, it was **amazing**! I had **sushi** at a special restaurant. And I also had **tuna**, **squid**, and **shrimp** while I was there. They were all really **delicious**. However, I did made a **mistake** at one point, I loaded up too much **wasabi**.
Yoko: **Wow!** that sounds like something I'd do.

セリフ内の太字の単語は，あらかじめ辞書に登録したもの（つまりテロップ表示するもの）である．現時点で登録した単語数は限られているが，今後辞書を拡張していく予定である．本システムでテロップ表示した画面例を図 4 に示す．

図 4. 画面表示例

4. 考察

4.1 技術面

基礎実験のアンケートにおいて，「表示に遅延を感じる」や「表示のタイムラグがある」といった意見があった．今回の機能強化では OpenCV の処理を高速化し，負荷を低減するために，映像の解像度やコーデック，画像のサイズなどの要素を配慮した．ただし，音声認識に使用されている Speech Recognition API は，リアルタイムで識別を行い，その結果を返すまでに時間が必要となる．そのため，画面表示と音声認識を合わせると遅延が生じることになる．SpeechRecognitionAPI では認識結果確定前の暫定認識結果も取得可能だが，確定結果よりも認識精度が低いという課題がある．

技術的には，辞書の強化が最大の課題である．現時点では一つの認識単語（フレーズ）に対して一つのテキスト，画像，フォント，フォントサイズだけを登録しているが，実際にはこれらはダイナミックに決められるものである．たとえば同じ単語を認識してもテロップを出す場合，出さない場合があるし，さまざまなテキストやイメージを表示する可能性もある．たとえば，「すごく楽しかった」という発話に対してテキストを表示する場合，「！」などを表示する場合，ハートマークのイメージを表示する場合などさまざまな場合があるであろう．

また，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．テロップ表示時間は 2 秒に設定しているが，テロップが連続して表示される場合，表示に問題が生じる可能性がある．この問題に対する対策として，そのときの顔画像の位置から一定の場所にテロップを表示するのではなく，たとえば顔のまわりを囲うようにテロップごとに少しずつ次のテロップ表示位置をずらす等の対策も考えらえる．

また，現在のシステムでは，画面内に複数の発話者が同時に存在する場合に，それぞれの話者を個別に識別して，表示を適切に仕分ける機能が実装されていない．このような機能を実現するためには，音声の特徴や波形の解析，発話者識別のためのアルゴリズムやモデルの導入が必要だと考える．特に，混合された音声の中から異なる話者を正確に識別するための音声分離技術や話者識別モデルの構築が求められる．

システムが遅延なくテロップを表示するための処理性能が重要であるが，特に音声認識を SpeechRecognition API 経由で行う場合に遅延が一定しないという課題もあり，この問題の対処も必要である．

4.2 ユーザー体験への影響の確認

表示の遅延が体験に与える影響の大きな一つの要因として，映像と音声の同期があり，遅延があると情報の理解や視聴体験に支障をきたす可能性がある．特に連続するテロップや音声との同期が取れない場合，情報が混乱することが懸念される．また，前述した話者を仕分ける機能は，ユーザーにとって，画面内の複数の発話者を正しく識別し，コミュニケーションや情報を理解するために重要な要素である．例えば，複数の人物が登場する動画では，発話者ではない人物の近くにテロップが出た場合，視聴者が混乱するようなことが考えらえる．これらを考慮し，表示時間あ表示位置に関して，良いユーザ体験を実現するためにはその程度の遅延が許容範囲であるかなども検討する必要があるだろう．

4.3 応用範囲

今回提案したテロップシステムは．既存の動画視聴だけでなくオンラインビデオ会議などへの活用も考えられる．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．現在，オンライン会議の効果や効率を上げるさまざまな手法が検討されているが [9]，発話にあわせてテロップ表示をすることで，やわらかなコミュニケーションの実現に貢献するであろう．

今回のテキスト，イメージ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．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可能であるが，どのような言い換えが可能であるかについては今後の検討課題である．最近急速に成長している生成 AI の利用等も候補となる．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る．

5. おわりに

本研究では，動画へのテロップ付加を自動的に行うためのシステムを提案し，システムの基本的な有効性確認と，一定の機能強化を行った．今後は，新たに構築したテロッ

プ表示システムの効果を検証するための評価実験を予定している．また，本稿ではエンタテインメントを目的としたテロップ付加について述べたが，このようなシステムの効果は，従来は字幕を用いていたようなケースで役立つ場合もあると考える．

謝辞 本研究は，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．

参考文献

[1] 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
[2] R. Hong, et al. Dynamic captioning: Video accessibility enhancement for hearing impairment, MM '10, pp. 421-430 (2010)
[3] A. Brown, et al. Dynamic subtitles: The user experience, TVX 2015, pp. 103-112 (2015)
[4] 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)
[5] Y. Hu, et al. Speaker-following video subtitles, ACM Transactions on Multimedia Computing, Communications and Applications, pp. 1-17 (2014)
[6] 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)
[7] R. Sasamoto: Impact caption as a highlighting device: Attempts at viewer manipulation on TV, Discourse, Context and Media Vol. 6, pp. 1-10 (2014)
[8] 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).
[9] 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)