Traverco AR: ARグラス上の字幕翻訳を用いた会話補助システム

西田　直人 *　　堀部　咲歩 †　　暦本　純一 ‡§

概要.　字幕翻訳を実時間にて提示し，母国語が違う者同士であっても，それぞれの母国語を用いて会話を行えるシステム Traverco AR を提案する．AR グラスを用いることにより，相手の表情やジェスチャなどのノンバーバルな動作を視認しながら会話を行うことができる．また，通訳者や音声通訳アプリといった audio-to-audio な通訳機構を介して会話を行う場合と異なり，発話にかかる時間分の遅延が発生しにくいと考えられる．さらに，本システムは文章媒体であるため，会話文の要約，または色分けといった提示手法を取り入れられることが考えられる．これにより，audio-to-audio な通訳機構に比べ，本システムを通じてより幅広いコミュニケーションのパターンが考えられる．

図 1. システムの概略図．日本語話者と英語話者の会話を想定している．日本語話者が日本語を話すと英語話者の AR グラス上に英訳された文章が表示され，逆に英語話者が英語を話すと日本語話者の AR グラス上に日本語訳された文章が表示される．

1　はじめに

現代において，世界中の人間とコミュニケーションを取るのは仕事や私生活において必須になっており，中には母国語が異なる人とコミュニケーションを行う場面も多く存在する．現在，人々は世界的な共通言語である英語を習得しコミュニケーションを取ることが多い．しかし、非母国語の習得には膨大な時間がかかるため、数多くの英語話者は言語の理解能力や発言能力が不十分なまま会話を行わざるを得ない．これに対し，通訳者，または通訳アプリなどの母国語が異なる人同士が会話を行うための補助が提案および実装されてきた．しかし，通訳者や通訳アプリといった audio-to-audio 型の通訳機構を介すると，発話にかかる時間分のタイムラグが生じ，伝達される情報量に対する所要時間は延びてしまう．また，audio-to-text 型の既存の実時間翻訳システムは持ち手が塞がる，および会話の際に相手の表情や動きが見えない問題がある．

よって，我々は AR グラス上に字幕翻訳を表示し，母国語が違う者同士の会話を補助するシステムを提案する（図 1）．実時間な字幕翻訳を表示することにより，母国語が違う者同士であっても，それぞれの母国語を用いて会話を行うことができる．さらに，AR グラスを用いることにより，相手の表情やジェスチャなどのノンバーバルな動作を確認しながら会話を行うことができる．

* 東京大学
† 東京大学
‡ 東京大学
§ Sony CSL Kyoto

2　関連研究

2.1　AI による通訳および翻訳

近年，使用言語が異なる人同士の会話を補助するための AI を用いたシステムは提案および実装されている [4][2]．しかし，audio-to-audio 型の通訳システムでは通訳結果を聞き取る必要があり，翻訳字幕を読むよりも会話のやり取りに時間がかかってしまう上，通訳音声を聞き逃す可能性もある [4]．

よって，本システムでは，audio-to-text 型の実時間翻訳システムによる字幕を提示することにより，audio-to-audio 型の通訳システムよりも速く安定的な情報伝達を実現することにした．

さらに，audio-to-text 型の既存の実時間翻訳システムは持ち手が塞がる，および会話の際に相手の表情や動きが見えない問題がある [2]．

よって，本システムでは，AR グラスを用いることにより，相手の表情やジェスチャなどのノンバーバルな動作を会話中に視認することを可能にした．

2.2 AR グラスへの字幕の表示

AR グラス上に字幕を表示する研究について，聾者や難聴者に向けたアクセシビリティ分野の文脈においてこれまで研究がなされてきた [3, 6]．これらの研究では，同一の母語話者を想定したシステムが多く，言語交換に利用されているものはない．Olwal らの研究においては，翻訳機能を活用した構想は提案されているが、現在の音声認識技術や翻訳技術を用いて実際に母国語同士で意思疎通がはかれるかは研究されていない [5]．

よって，本研究では，翻訳字幕を AR グラス上に投影したシステムにより異なる母語話者同士が意思疎通を図れるか，およびシステムがユーザ体験に与える影響を調べる．

3 実装

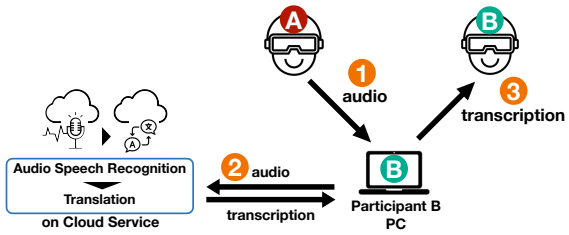

図 2. ユーザ A の発話の音声データを翻訳する際のワークフロー．1）ユーザ A の音声データがユーザ B の PC に入力される．2）音声データが音声認識および翻訳される．3）字幕翻訳がユーザ B の AR グラス上に表示される．

我々は，音声を AR グラス使用者の母国語に翻訳し，字幕として提示するシステムを実装した（図 2）．

実装言語に C#を用いた．AR グラスに投影する空間を作成するために Unity を用いた．音声認識および翻訳のために Microsoft Azure Cognitive Service API を用いた．字幕翻訳の保存のために firebase を用いた．AR グラスおよび PC について，Magic Leap 2 および MacBook Pro 13 inch Apple Silicon 版を用いた [1]．

ユーザ A がマイクから音声データをユーザ B の PC へ入力すると，ユーザ B の PC はクラウド上の音声認識および翻訳を行う API に接続し，音声認識および翻訳を行う．API から翻訳された文章が返された後，ユーザ B の AR グラスには翻訳された文章（翻訳字幕）が表示される．翻訳については，あらかじめ翻訳元の言語および翻訳先の言語をプログラム内にて指定する．

4 パイロットスタディ

本システムを用いて非母国語話者と話した際の体験を調査するために，パイロットスタディを実施した．筆者と同じ研究室内の学生 3 名，　研究室外の教員および学生 2 名（1 名中国人，1 名ニュージーランド人，1 名ドイツ人，2 名日本人．4 名男性）が体験した．得られたフィードバックとしては以下のものがあった：

- 周囲で話している声も認識されている気がする
- 翻訳元の言語を変更するためにプログラムを書き換えないといけないのは不便である
- プレゼンなどの長い文章を聞いていると，提示される文章量が長くなる
- 入力音声について，口語と文語では難易度が異なる可能性がある
- 文法が異なる言語と似ている言語では難易度が異なる可能性がある
- 目線といった他の入力も用いると更なる機能が期待できる
- 多人数における会話も体験したい

5 議論および今後の展望

議論としては 3 点ある．

まず，本システムでは音声入力を全指向性マイクから行っていたため，AR グラス使用者の声も入ってしまい翻訳字幕が乱れることが分かった．よって，システムの翻訳精度向上のために，単指向性マイクを用いる，または音源分離を行う必要がある．

また，現在の実装では，翻訳元および翻訳先の言語をプログラム内からあらかじめ指定する必要がある．本課題に対し，音声認識 API に接続する前に言語検出を行えば翻訳元の言語を指定する必要がなくなる可能性がある．また，AR システム上にて翻訳元言語を選ぶ機能をつけることにより，ユーザの負担が軽減される可能性もある．

さらに，本システムは同じ言語の翻訳にも用いることができる可能性がある．例として，英語話者同士でも訛りによってはお互いの発言内容を聞き取りづらいという場面が考えられる．この場合，本システムの音声認識機能のみを使用することにより，訛りを字幕提示することにより意思疎通の支援を行うことができる．他の例としては，第二外国語学習者の文法的誤りを含んだ発話内容が母語話者に理解されない場面も考えられる．この場合，本システムの音声認識機能に加え，文法誤り訂正を行った文章を提示することにより，意思疎通の支援を行える．

今後の展望としては，audio-to-audio 型の翻訳システムと比較した際の情報伝達速度の比較や，既存の audio-to-text 型の翻訳システムと比較した際に，AR グラスを用いることにより本システムがユーザ体験に及ぼす影響を調べ，さらに，ユーザ体験を向上させるために適切な字幕提示手法を模索していく予定である．

謝辞

本研究は JST ムーンショット型研究開発事業グラント番号 JPMJMS2012, JST CREST グラント番号 JPMJCR17A3, および東京大学ヒューマンオーグメンテーション社会連携講座の支援を受けたものです。

参考文献

[1] Magic Leap 2. https://www.magicleap.com/magic-leap-2（最終閲覧日：2022 年 11 月 23 日）.

[2] Poketalk. https://pocketalk.jp/（最終閲覧日：2022 年 11 月 23 日）.

[3] D. Jain, B. Chinh, L. Findlater, R. Kushalnagar, and J. Froehlich. Exploring Augmented Reality Approaches to Real-Time Captioning: A Preliminary Autoethnographic Study. DIS '18 Companion, pp. 7–11. ACM, 2018.

[4] Meta. Using AI to Translate Speech For a Primarily Oral Language, 2022. https://about.fb.com/news/2022/10/hokkien-ai-speech-translation/（最終閲覧日：2022 年 11 月 23 日）.

[5] A. Olwal, K. Balke, D. Votintcev, T. Starner, P. Conn, B. Chinh, and B. Corda. Wearable Subtitles: Augmenting Spoken Communication with Lightweight Eyewear for All-Day Captioning. UIST '20, pp. 1108–1120. ACM, 2020.

[6] Y.-H. Peng, M.-W. Hsi, P. Taele, T.-Y. Lin, P.-E. Lai, L. Hsu, T.-c. Chen, T.-Y. Wu, Y.-A. Chen, H.-H. Tang, and M. Y. Chen. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. CHI '18, pp. 1–10. ACM, 2018.