画像を中間言語とした多言語音声翻訳システム

趙　澤毅†　　林　　実†
† 明星大学理工学研究科

1. はじめに

伝統的な多言語音声翻訳は一つの言語から別の言語への直接翻訳を行い、言語の増加に伴うモジュール増の問題がある。それらモジュール数を減らすために、これまで中間言語を用いた手法が研究されてきた[1]。中間言語は、異なる2言語間の直接翻訳ではなく、中間言語を介して翻訳を行う。しかしながら中間言語の翻訳過程において、初段回の翻訳誤りが第二段回の翻訳誤りを増幅させる問題点がある[2]。

そこで本研究では、画像を中間言語として多言語音声翻訳システムの構築方法を提案する。

2. 中間言語

図1は多言語音声翻訳の相互依存性を双方向矢印で示し、図 1(a) は伝統的な翻訳を、図 1(b) は中間言語「X」を用いモジュール数を減らす翻訳をそれぞれ示している。

図 1. (a)伝統的な翻訳　　(b)中間言語を用いた翻訳

2. 1 画像による中間表現

本研究では中間言語の代わりに画像を用いる方法を提案する。画像は多くの情報を提供し、その視覚的な表現は翻訳の初段回誤りと誤解釈を減らすことができる。

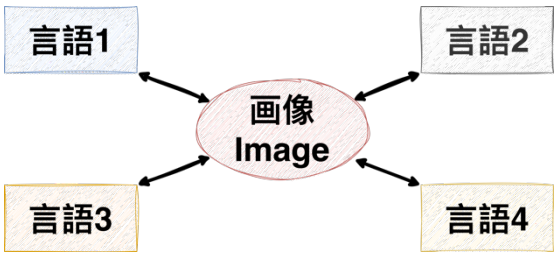

図 2. 画像による中間表現

3. 本提案システム

実験における画像を中間言語とした多言語音声翻訳システムの構成を図 3 に示す。多言語音声翻訳の流れは、まず音声をアップロードし、Whisper モデルに送り、モデルは自動的に言語を識別し、テキストを出力する。次に DALL-E モデルを用い音声認識で得られたテキストから対応する画像を生成する。生成された画像は言語の中間表現として用いる。更に設定した対象言語を生成画像からコンピュータビジョン技術を用い対応する画像キャプションを生成し、対象言語の翻訳結果として用いる。最後に生成された画像キャプションは Text-to-Speech により対象言語の音声を生成する。

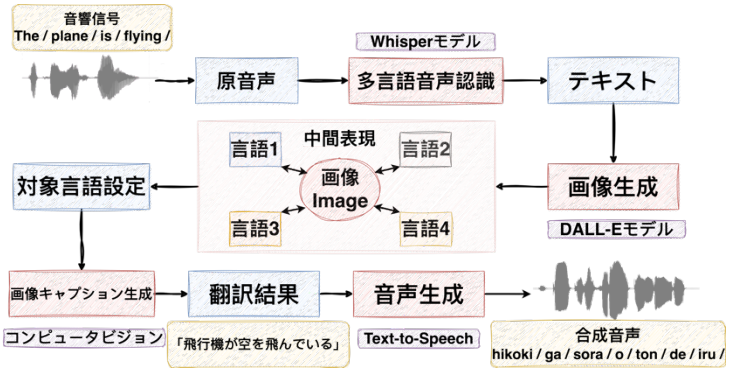

図 3. 本提案システムの構成図

4. 実験方法

本実験は、英語、中国語、日本語、およびスペイン語を対象言語とし、次に示す語彙組合せで実験を行った。

A: 冠詞+形容詞+名詞
B: 冠詞+名詞+動詞
C: 名詞+介詞(Prep.)+名詞+副詞

これらの組合せを用い、1 回、5 回、および 10 回の生成回数を設定し、中間表現の画像による翻訳の出力と参照訳として DeepL の出力を比較し、BLEU 値により実験の評価を行った。

5. 実験結果

表 1. 4言語の3種語彙組合せの平均 BLEU 値

	A	B	C
1 回	0.3912	0.3331	0.3355
5 回	0.3869	0.3333	0.5073
10 回	0.4112	0.3641	0.4716

6. 考察

表 1 から分かるように、全単語の組合せに対し、生成回数が増えるにつれて翻訳精度が向上する。今後、画像生成困難な単語組合せの改善および多言語モデル[3]の改良により、方言など多様な言語翻訳への拡張を目指し、画像を中間言語としての翻訳精度を高めて行く。

7. まとめ

本研究において提案された「画像を中間言語とした多言語音声翻訳システム」は、その革新的なアプローチにより、多言語数増に伴う翻訳モジュール数の増大という課題に根本的な解決策をもたらすものと期待される。

参考文献

[1] Adusumilli, K. K. 2007. "Natural languages translation using an intermediate language. "
[2] Gispert, A., 2006. "Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish."
[3] Yang, Zhengyuan, et al. "The dawn of lmms: Preliminary explorations with gpt-4v (ision)."