INTRODUCTION

• Speech-to-Speech Translation (S2ST)
  - Enhance multilingual communication
  - Relies on a massive amount of parallel source-target speech data
  - Parallel data is often unavailable
• Human Infants
  - Multilingual acquisition ability
  - Allow them to acquire languages based on visual information

The paper proposes VGSAlign:
- Attempt to mimic human infants’ behavior
- Aim to discover the speech pairs data for S2ST
- Find speech similarity of source and target languages based on corresponding visual context
- Utilize self-supervised visually grounded speech model
- Unable to deal with S2ST for unknown, unpaired, untranscribed languages

VGSALIGN FRAMEWORK

The system combines two modules:

1. Image-Image Similarity Module
2. Cross Speech-Image Similarity Module

Leveraging self-supervised visually grounded speech models as encoders for image and audio

Figure 1: Bilingual speech alignment by visual-based information.
(Note: The image is from the MS-COCO dataset)

Figure 2: The overview of the VGSAlign framework.

CONCLUSIONS

- VGSAlign can be applied to any other languages
- Allow mapping speech from the source to the target languages
- Able to determine whether two given speech in different languages are semantically paired without the need for text and knowledge about the language

FUTURE DIRECTIONS

- Perform speech-to-speech translation for unknown, unpaired, untranscribed languages by using data from the VGSAlign system
- Investigate the obtained speech-image co-embeddings in order to get pseudo-speech-speech pairs

REFERENCES