LIDILEM

UNIVERSITÉ Grenoble Alpes

Models of visually grounded speech signal pay attention to nouns: a cross-linguistic experiment on English and Japanese

William N. Havard, Jean-Pierre Chevrot, Laurent Besacier
{william.havard,jean-pierre.chevrot,laurent.besacier}@univ-grenoble-alpes.fr

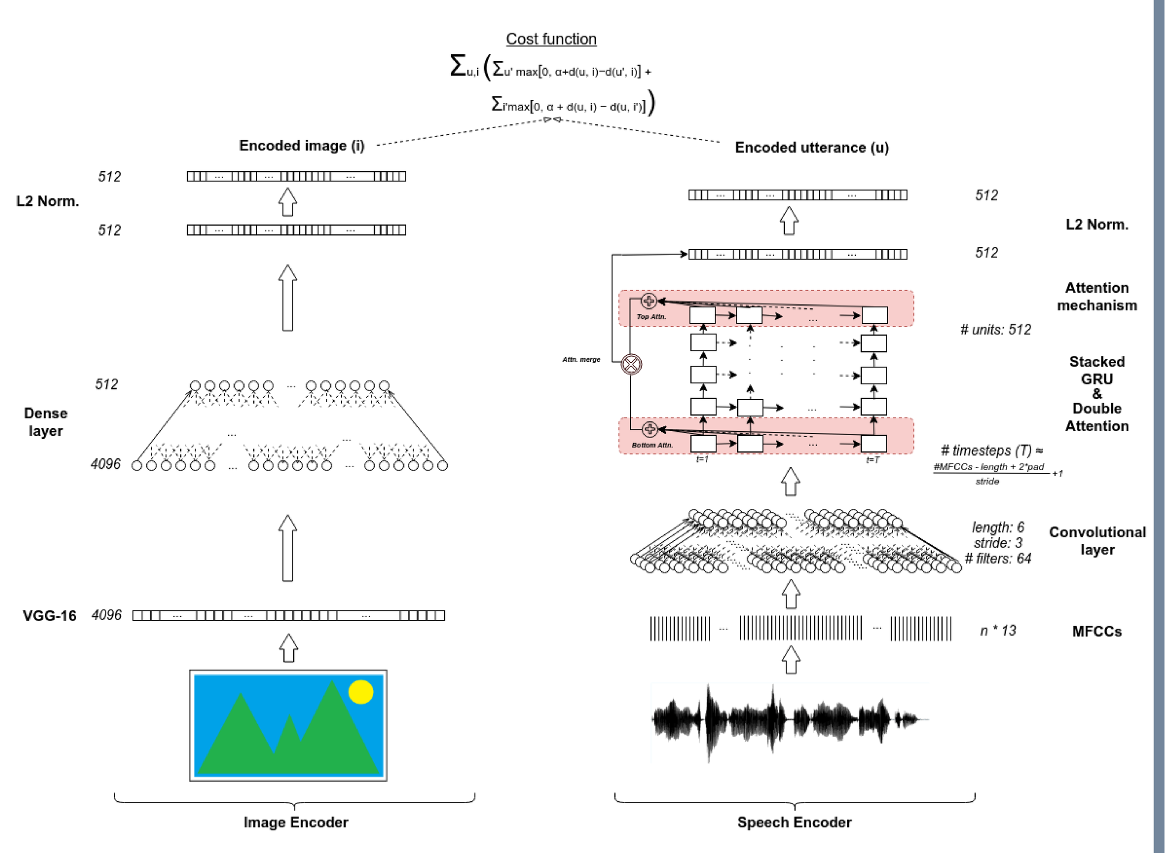
LIG

GETALP

1. Introduction

- We investigate the **behaviour of attention** in neural models of visually grounded speech trained **on two languages: English and Japanese.**
- Experimental results show that **attention focuses on nouns** and this behaviour holds true for two very typologically different languages. We also draw **parallels between artificial neural attention and human attention** and show that **neural attention focuses on word endings** as it has been theorised for human attention.
- Finally, we investigate how **two visually grounded monolingual models** can be used to perform **cross-lingual speech-to-speech retrieval**.
- For both languages, the enriched bilingual (speech-image) corpora with POS tags and forced alignments **are distributed** to the community.

2. Models



Model	R@1	R@5	R@10	$\tilde{r}$
English	0.060	0.195	0.301	25
Japanese	0.054	0.180	0.283	28

- Architecture based on [1]
- Two attention mechanisms
- Projects an image and its spoken description in a **common representation space**
- One model for each language: **English** and **Japanese**

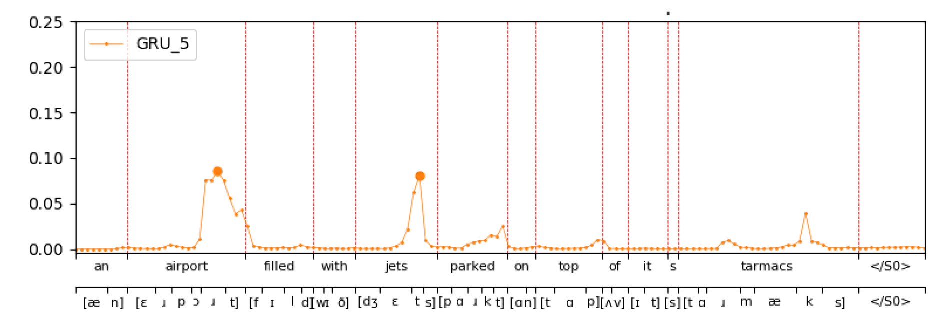
3. Data

- Comparable corpora featuring the same images:
  - **MSCOCO** [2] for **English**
  - **STAIR** [3] for **Japanese**
- Set of **images** paired to **5** human-written **captions**
- Google TTS to generate **synthetic speech** for English and Japanese
- Data and metadata available here: https://github.com/William-N-Havard/VGS-dataset-metadata

References

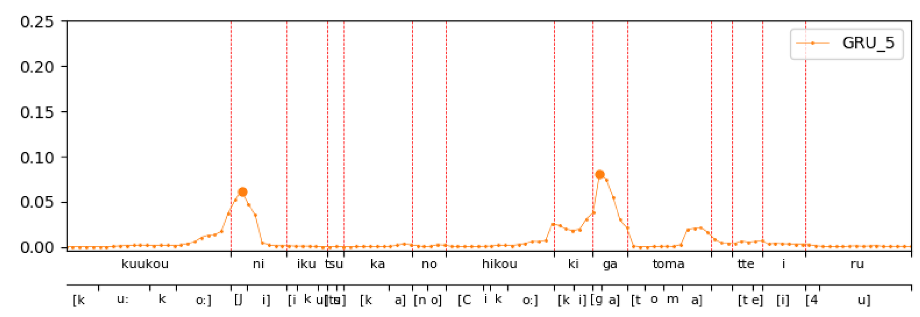
[1] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *ACL*, 2017.
[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
[3] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. Stair captions: Constructing a large-scale japanese image caption dataset. In *ACL*, 2017.
[4] David Harwath, Galen Chuang, and James R. Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *ICASSP*, 2018.
[5] C.A. Ferguson and D.I. Slobin. *Studies of child language development*. New York : Holt, Rinehart and Winston, 1973.

4. Attention



English caption

Common picture

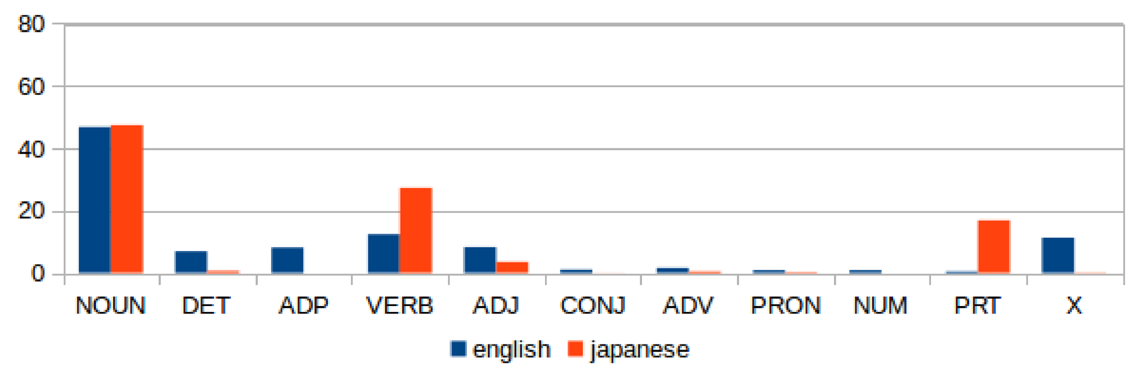


Japanese caption

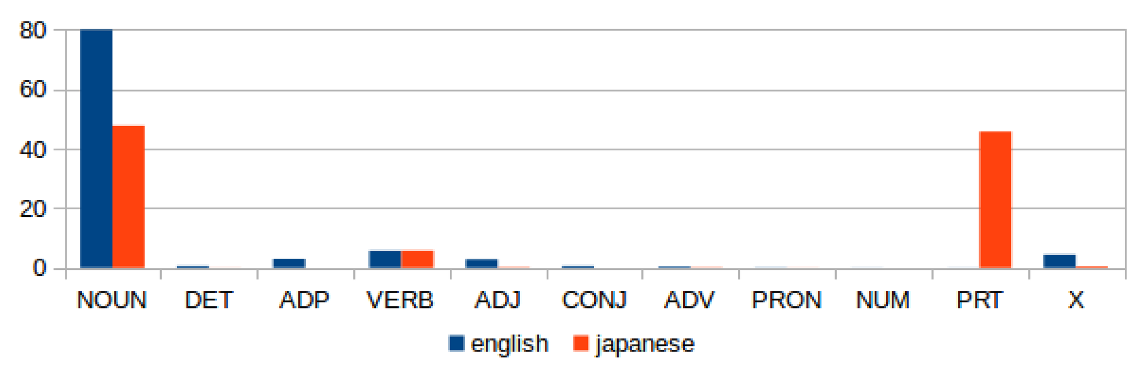
("Several planes are stopped at the airport")

- Extraction of the attention weights for the **English** and **Japanese** captions
- Automatic peak detection
- Statistics on Part-of-Speech (POS) distribution beneath peaks

5. What do models pay attention to?



Baseline POS distribution if attention peaks were to occur randomly



POS distribution of words under detected attention peaks

- **English**
  - **82%** of the peaks are located **above nouns**. Far above **corpus frequency** which is **47%**
- **Japanese**
  - **47.79%** above **nouns**
  - **Language-specific behaviour**: **45.77%** of the peaks above **particles**
- **Child language acquisition** and **noun bias**: children learn nouns before any other category **Japanese children** rely on the "GA" particle for word segmentation

English			Japanese			
word	peak freq.	ref. freq	word	gloss	peak freq.	ref. freq
toilet	2.16	0.17	ga	subject part.	17.83	5.25
baseball	1.84	0.22	no	topic part.	9.53	6.24
train	1.71	0.25	o	direct object part.	6.6	0.59
giraffe	1.6	0.11	ni	location part.	6.55	3.58
skateboard	1.57	0.14	de	location part.	1.81	1.72

6. Towards Speech-to-Speech Retrieval

- Speech-to-Speech retrieval using **images as pivots** with **two monolingual models**
  - 2 monolingual models (EN & JP) trained on the non-overlapping halves of the train set
  - For each speech utterance query in source language u_{src}, find nearest speech utterance in target language u_{tgt} which minimises the cumulated distance $d(u_{src}, i) + d(i, u_{tgt})$ among all pivot images i.
  - Evaluated on a subset of 1k captions. Given a speech query in language *src* which we know is paired with image I, we assess the ability of our approach to rank the matching spoken caption in language *tgt* paired with image I in the top 1, 5, and 10 results.

Query	R@1	R@5	R@10	$\tilde{r}$
EN → JP	0.087	0.327	0.519	9.94
JP → EN	0.087	0.326	0.521	9.84
[4] EN → HI	0.034	0.114	0.182	–
[4] HI → EN	0.033	0.121	0.203	–

EN	This is a display of donuts on a couple shelves
JA	いろいろな種類のドーナツが並べられている
Trans.	Different kinds of donuts are lined up
EN	A living room with some brick walls and a fireplace
JA	ソファーやテーブルや暖炉のある西洋風の部屋
Trans.	Western-style room with sofa, table and fireplace

7. Conclusion

- Attention in a neural model of visually grounded speech mainly focuses on **nouns as children do**
- This behaviour holds **true for two very typologically different languages** such as **English** and **Japanese**
- Attention develops **language-specifc mechanisms** to detect relevant information
- **Future work:** explore the behaviour of a Japanese-English bilingual model