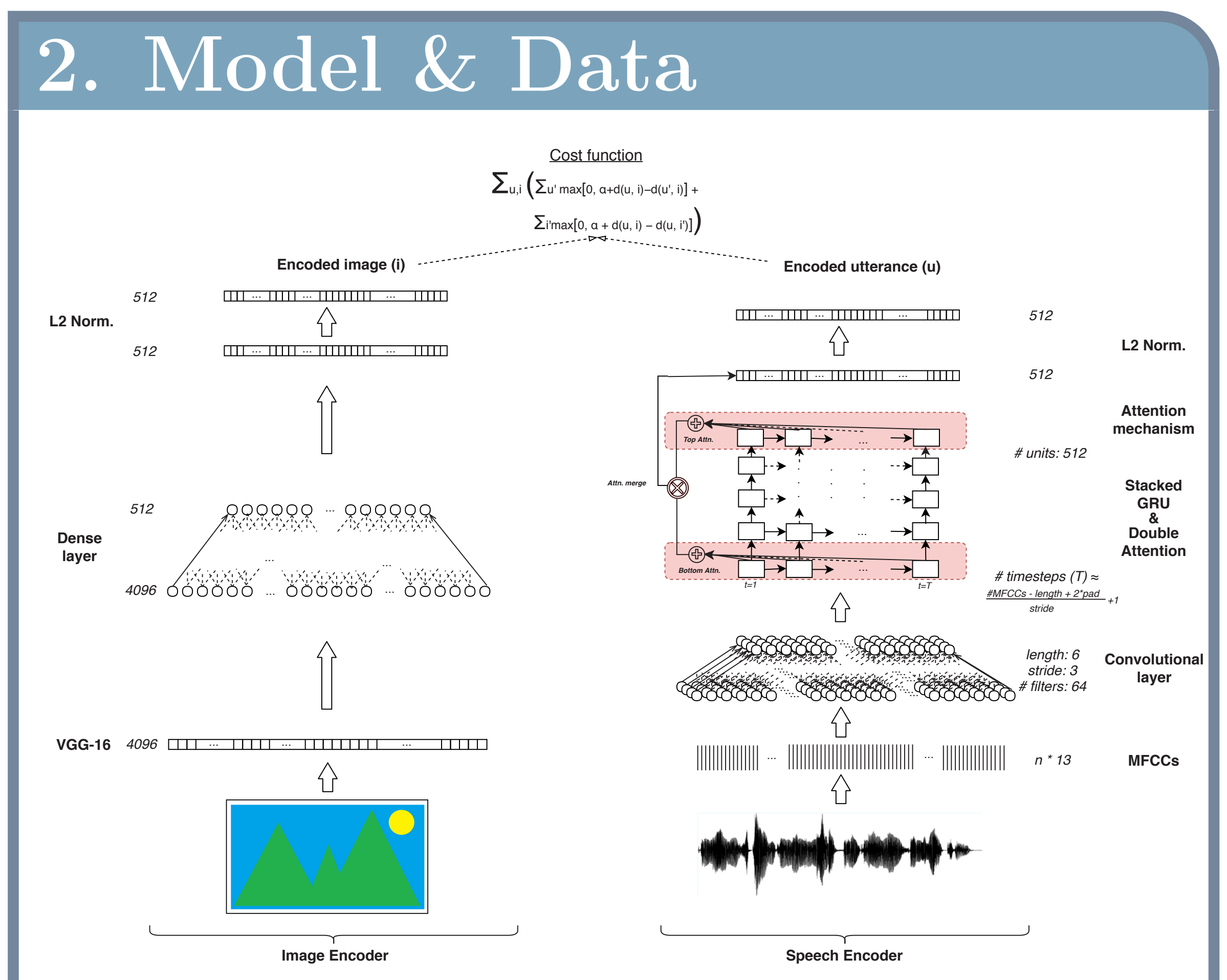
Word Recognition, Competition, and Activation in a Model of Visually Grounded Speech

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

1. Introduction

- We investigate if a neural model of visually grounded speech (VGS) is able to **map isolated words** to their **visual referents** despite having been **trained on full utterances** (*word recognition*), **how** these words are being **activated** (*word activation*), and if **multiple words** are **simultaneously activated** (*word competition*).
- We introduce a methodology stemming from linguistics – the **gating paradigm** – to **analyse the representations** learnt by a VGS model. This methodology could also be used to **analyse** the representation of **any neural model** handling **speech**.

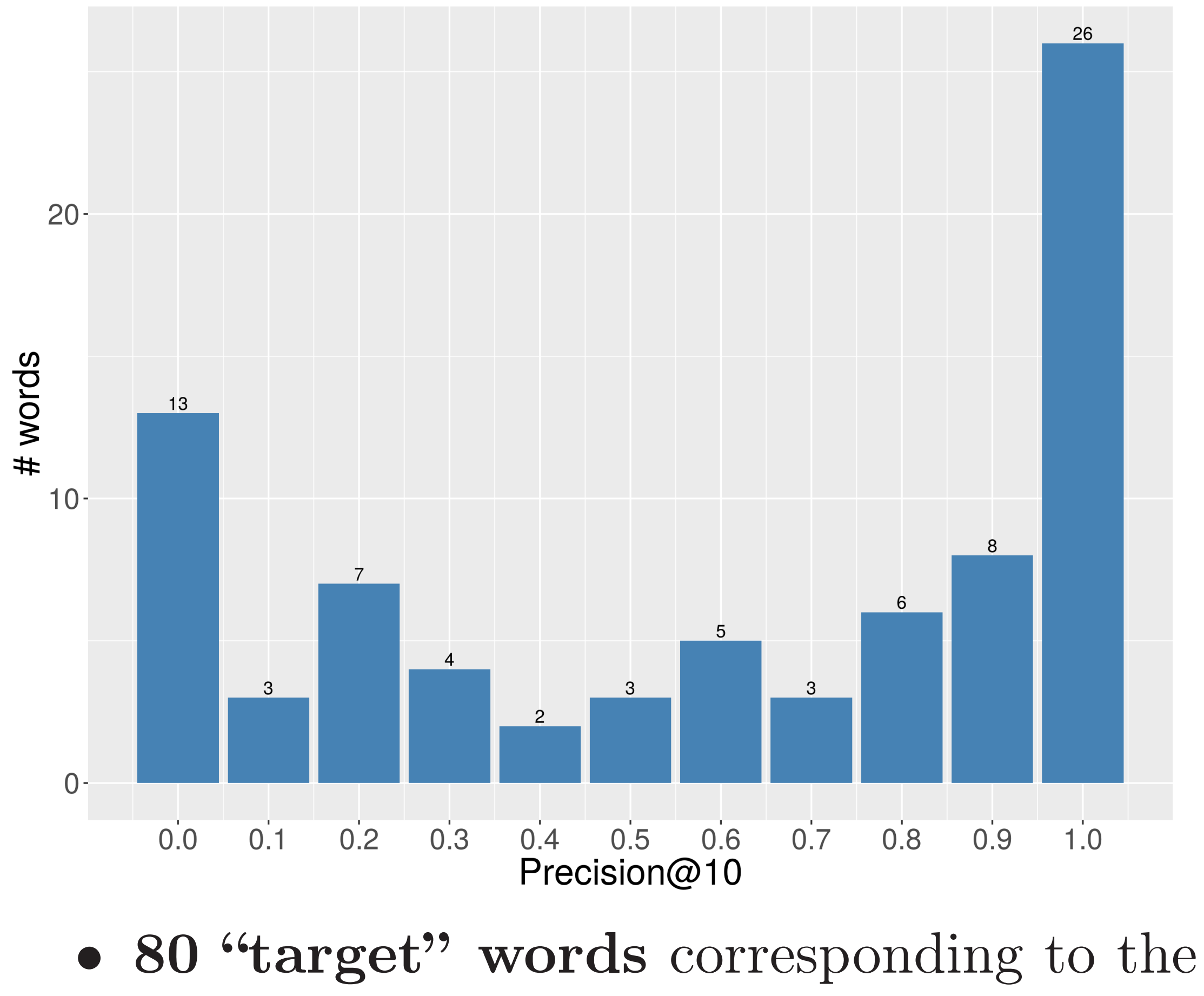
2. Model & Data



- Architecture based on [1]
- Projects an image and its spoken description in a **common representation space**

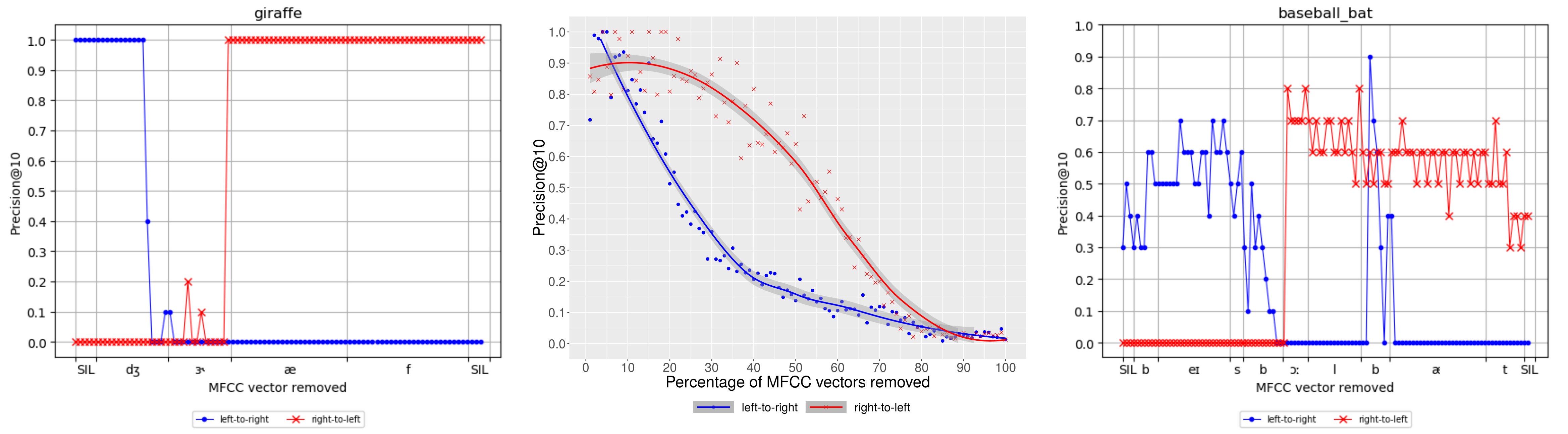
- Synthetically Spoken **MSCOCO** [1, 2]
- Set of 113k **images** paired to **5** spoken captions

3. Word Recognition



- **80 "target" words** corresponding to the **80 object classes** in MSCOCO
- Model is able to **map isolated words** to their **visual referents**
- **Not all the words** are equally well recognised
- Concepts corresponding to **frequent words** as well as **bigger objects** are **better recognised**

4. Word Activation



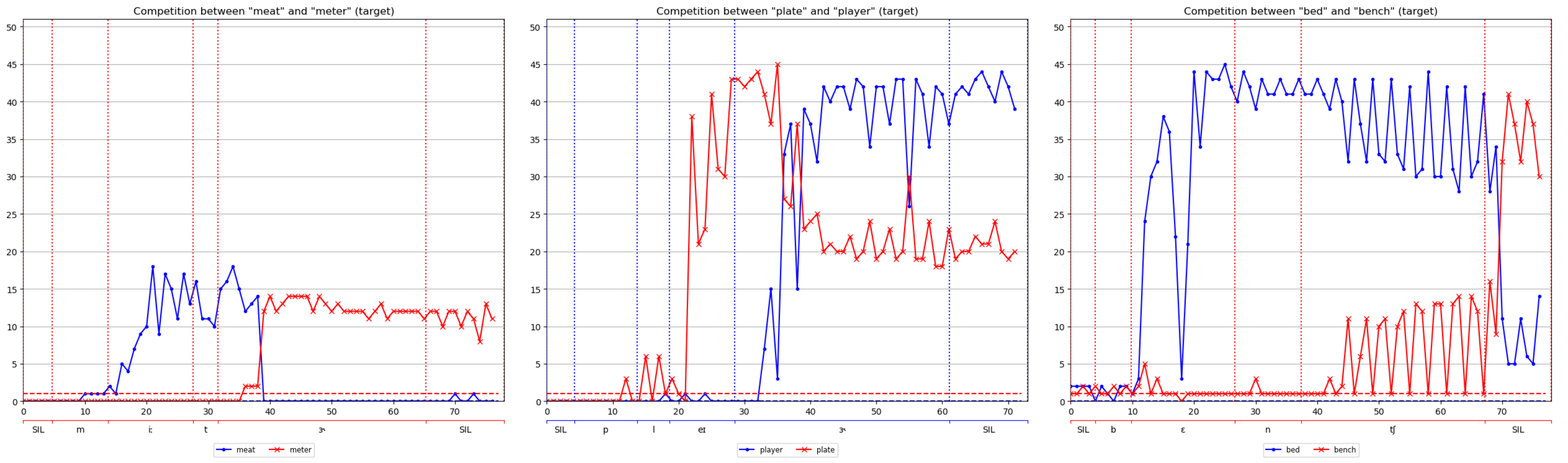
(a) Precision@10 for the word "giraffe"

(b) Precision@10 averaged over the 80 target words

(c) Precision@10 for the word "baseball bat"

- COHORT **model** [3] stipulates that **word onsets** play a **crucial role** in word recognition
- **Gating paradigm** [4]: neural model is fed with **truncated version of a target word**, each truncated version comprising a larger part of the target word
- Model is **robust to truncation when it is carried out right-to-left** but **not** when it is carried out **left-to-right**: network very **sensitive to word onsets**
  - Model **fails to retrieve** pictures of giraffes when **first phoneme /ʤ/ is removed** and only /ɝæf/ is left
- Gating enables us to understand the **internalised pseudo-words** by the network
  - /ʤɝ/ is enough to activate the representation of the word "giraffe"
  - Both **"baseball bat"** and **"bat"** are mapped to the **same referent**

5. Word Competition



(a)

(b)

(c)

- According to the COHORT **model** [3]:
  - 1^{st} phoneme of a word activates **all the words starting** with the **same phoneme**
  - Words **"deactivate"** when speech input becomes **inconsistent with internalised representation**
- **Competition**: words **compete to stay activated** even though the input only partly matches the internalised representation
- No initial **cohort**: **words** are **activated sequentially** (fig. *a*) and not simultaneously
- Some words **remain highly activated** (as in fig. *b*) even though the input is inconsistent with the target word

6. Conclusion

- A neural model of VGS is robust to isolated word stimuli suggesting an **implicit segmentation** into **sub-units**.
- Our model needs to have access to the **first phoneme of a word** to activate its representation.
- The **beginning of a word** is enough to activate the representation of a given concept (e.g. /ʤɝ/ for "giraffe").
- Our model activates **representations sequentially** and **not simultaneously**.
- We used the **gating paradigm** [4] to analyse the representation learnt by our model that could also be applied to **understand ASR systems**.

[1] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622. Association for Computational Linguistics, 2017.
[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, T. Pajdla, B. Schiele, and T. Tuytelaars. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.
[3] William D. Marslen-Wilson. Functional parallelism in spoken word-recognition. *Cognition*, 25(1):71 – 102, 1987. Special Issue Spoken Word Recognition.
[4] François Grosjean. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28(4):267–283, Jul 1980.