Models of Visually Grounded Speech Signal Pay Attention to Nouns: A Bilingual Experiment on English and Japanese

William N Havard, Jean-Pierre Chevrot, Laurent Besacier

MODELS OF VISUALLY GROUNDED SPEECH SIGNAL PAY ATTENTION TO NOUNS: A BILINGUAL EXPERIMENT ON ENGLISH AND JAPANESE

William N. Havard[1][2], *Jean-Pierre Chevrot*[2], *Laurent Besacier*[1]

[1]LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, 38000 Grenoble, France
[2]LIDILEM, Univ. Grenoble Alpes, 38000 Grenoble, France

ABSTRACT

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 part-of-speech tags and forced alignments are distributed to the community for reproducible research.

Index Terms— grounded language learning, attention mechanism, cross-lingual speech retrieval, recurrent neural networks.

1. INTRODUCTION

Over the past few years, there has been an increasing interest in research gathering the Language and Vision (LaVi) communities. Multimodal corpora such as Flickr30k [1] or MSCOCO [2] containing images along with natural language captions were made available for research. They were soon extended with speech modality: speech recordings for the captions of Flickr8k were collected by [3] *via* crowdsourcing; spoken captions for MSCOCO were generated using Google Text-To-Speech (TTS) by [4] and using Voxygen TTS by [5]; extensions of these corpora to other languages than English, such as Japanese, were also introduced by [6]. These corpora, as well as deep learning models, lead to contributions in multilingual language grounding and learning of shared and multimodal representations with neural networks [4, 7, 8, 9, 10, 11, 12, 13].

This paper focuses on computational models of visually grounded speech that were introduced by [14, 4]. Learned representations of such models were analyzed by [11, 7, 4]: [11] introduced novel methods for interpreting the activation patterns of recurrent neural networks (RNN) in a model of visually grounded meaning representation from textual and visual input and showed that RNN pay attention to word tokens belonging to specific lexical categories. [4] found that final layers tend to encode semantic information whereas lower layers tend to encode form-related information. [7] showed that a non trivial amount of phonological information is preserved in higher layers, and suggested that the attention layer focuses on semantic information.

Such computational models can be used to emulate child language acquisition and could shed light on the inner cognitive processes at work in humans as suggested by [15]. While [11, 7, 4] focused on analyzing speech representations learnt by speech-image neural models from a phonological and semantic point of view, the present work focuses on lexical acquisition and the way speech utterances are segmented into lexical units and processed by a computational model of visually grounded speech. We analyze a key component of the neural model – the attention mechanism – and we observe its behaviour and draw parallels between artificial neural attention and human attention. Attention indeed plays a key role in human perceptual learning, as stated by [16].

Contributions. We enrich an existing speech-image corpus in English with forced alignments and part-of-speech (POS) tags and analyse which parts of the spoken utterances the neural model attends to. In order to put these experiments in a cross-lingual perspective, we also experiment on a similar corpus in Japanese.[1] We show that the attention mechanism mostly focuses on nouns for both languages. We also show that our Japanese model developed a language-specific behaviour to detect relevant information by paying attention to particles, as Japanese toddlers do. Moreover, the bilingual corpus allows us to demonstrate that images can be used as pivots to automatically align spoken utterances in two different languages (English and Japanese) without using any transcripts. This preliminary result, in line with previous findings of [8], confirms that neural speech-image models can capture a cross-lingual semantic signal, a first step in the perspective of learning speech-to-speech translation systems without text supervision.

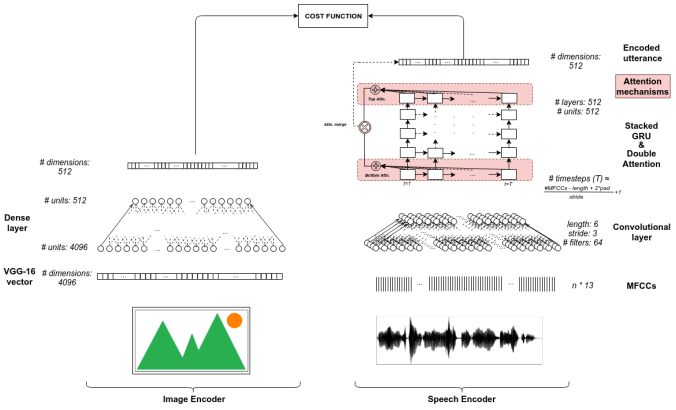


Fig. 1: Neural model of visually grounded speech used in our experiments.

This work was supported by grants from NeuroCoG IDEX UGA as part of of the "Investissements d'avenir" program (ANR-15-IDEX-02)

[1]Both enriched corpora are available on https://github.com/William-N-Havard/VGS-dataset-metadata.

(a) (b) (c)

Fig. 2: Attention weights over an English (2a) and Japanese caption (2c), both describing the same picture (2b). Attention peaks in the English caption are located above "AIRPORT" and "JETS". Attention peaks in the Japanese caption are located above "NI" (particle indicating location) and "GA" (particule indicating the subject of the sentence). Red dotted lines show token boundaries. Large orange markers show automatically detected peaks. Japanese caption reads: "Several planes are stopped at the airport"

2. MODEL OF VISUALLY GROUNDED SPEECH

The model we use for our experiments is based on that of [4]. It is trained to solve an image retrieval task: given a spoken description it retrieves the closest image that matches the description. To do so, the model projects an image and its spoken description in a common representation space, so that matching image/utterance pairs lie near while mismatching image/utterance pairs lie apart.

2.1. General architecture

The model (see figure 1) has two components: an image encoder, and a speech encoder. At training time, the network is presented with images and their corresponding spoken descriptions and tries to minimise the following loss function:

$$\sum_{u,i}\left(\sum_{u'}\max[0,\alpha+d(u,i)-d(u',i)] + \sum_{i'}\max[0,\alpha+d(u,i)-d(u,i')]\right) \quad (1)$$

This loss function encourages the network to minimise by a margin α the distance $d(u,i)$ between the encoded image i and the encoded utterance u belonging to matching image/utterance pairs while making the distance greater for mismatching image/utterance pairs.

2.2. Encoders

The image encoder takes VGG-16 ([17]) pre-calculated vectors as input[2] instead of raw images. It only consists of a dense layer that learns how to shrink the 4096 dimensional VGG-16 input vector to a 512 dimensional vector, which is then L2 normalised. The speech encoder (input is 13 MFCC vectors instead of raw speech) consists of a convolutional layer followed by 5 stacked recurrent layers. Contrary to the original model ([4]), we used GRU units instead of RHN units.[3] Results are still acceptable (see Table 1) even if GRU architecture scores worse than original RHN one.

2.3. Attention mechanism

One of the key component of the model is its attention mechanism. The model computes a weighted sum of the GRU activations at all timesteps as following: $\sum_t \alpha_t h_t$. Knowing by how much a given vector has been weighted gives us an insight on which portions of the speech signal the network relies to make its predictions (see Figure 2). In the original architecture ([4]), attention follows the last recurrent layer. To have more insight on the representation learnt by the network, we added an attention mechanism after the first recurrent layer. Final vector produced by the speech encoder is a dot product of the vectors produced by both attentions. However, for the sake of clarity, we will only report in this paper results on the attention weights of the top attention mechanism GRU5 (after the fifth recurrent layer).[4]

3. ENGLISH AND JAPANESE CORPORA

The corpora we use for our experiments are based on MSCOCO [2]. MSCOCO is a dataset initially thought for computer vision purposes, mainly automatic image captioning. The dataset consists of a set of images, each paired with 5 written captions describing the image. All captions were written in English by humans and faithfully describe the content of the image. The Japanese corpus we use is based on the newly created STAIR dataset [6]. Using the same methodology as [2], [6] collected 5 Japanese captions for each image of the original MSCOCO dataset. As for the original MSCOCO dataset, Japanese captions were written by native Japanese speakers. It is worth insisting on the fact that these Japanese captions are original captions and not plain translations of their English equivalents. MSCOCO and STAIR are thus comparable corpora. We trained our model on extended versions of MSCOCO and STAIR. Spoken COCO dataset was introduced by [4] for English. We followed the same methodology as [4] and generated synthetic speech for each caption in the Japanese STAIR dataset. We created the spoken STAIR dataset so it would follow the exact same train/val/test[5] split as [4]. We thus have two comparable corpora: one featuring images and spoken captions in English, and another one featuring the same images and spoken captions in Japanese. This allowed us to compare the behaviour of the same architecture on two typologically different languages.

We forced aligned each spoken caption to its transcription (using the Montreal Forced Aligner [18] and Maus Forced Aligner [19] for English and Japanese respectively), resulting in alignments at word and phone level. We also tagged each dataset using TreeTagger [20] for English and KyTea [21] for Japanese. As the tagset of both taggers differs, we mapped each POS to its Universal POS equivalent [22] enabling us to compare the POS distribution of each corpus.[6]

[2] VGG networks are trained to label images with a set of 1000 object categories from ImageNet.

[3] In fact, we aim at having a simpler model whose internal representations would be easier to understand as we also intend to study the gating mechanism in the future.

[4] Adding a second attention mechanism improves our results by -3 $\widetilde{r}$.

[5] 566 435, 25 000, and 25 000 captions in each set respectively.

[6] We decided to map KyTea's TAIL tags – word conjugation – to Univer-

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

Table 1: Recall at 1, 5, and 10 results as well as median rank $\widetilde{r}$ on a speech-image retrieval task (test part of our datasets with 5k images). Original implementation by [4] with RHN reports median rank $\widetilde{r} = 13$ on English dataset. Chance for median rank $\widetilde{r}$ is 2500.5.

4. WHAT DO MODELS PAY ATTENTION TO?

We first train two monolingual models for English and Japanese on the train set (566 435 spoken captions) of the corpora for 15 epochs. Baseline results are similar for English and Japanese (see Table 1).

To analyse the behaviour of the attention mechanism of our model, we encoded each caption of the test set and extracted the attention weights α_t, resulting in an array of t weights. We then used a peak detection algorithm[7] to detect local maxima in the attention weights and thus know which timesteps were given the highest weights (large orange markers in Fig. 2). We only considered peaks that were at least 60% as high as the highest detected peak in the utterance.

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
sign	1.33	0.19	piza	"pizza"	1.47	0.13
kitchen	1.17	0.18	to	"with" part.	1.04	1.37
with	1.13	2.09	ke:ki	"cake"	1.02	0.1
frisbee	1.11	0.11	shimauma	"zebra"	0.99	0.09
cake	1.03	0.11	suke:tobo:do	"skateboard"	0.98	0.13

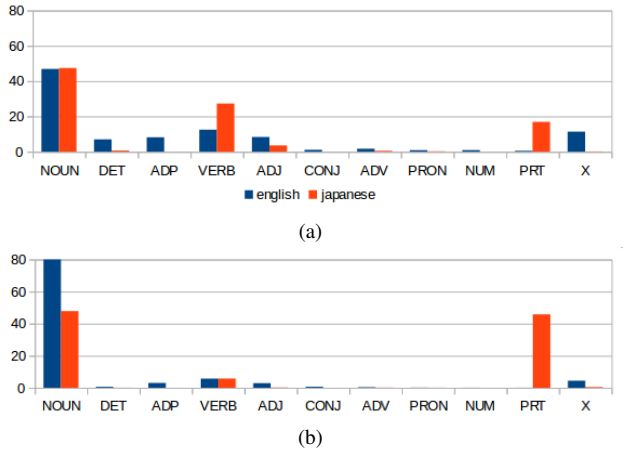
Table 2: Top 10 focused words for English and Japanese. "Peak freq." refers the number of attention peaks (in %) above a given word. "ref. freq." refers to the frequency of the same word token in the training set.

4.1. Which morpho-syntactic categories are highlighted by attention?

Having a timestep aligned speech signal for each language enables us to see above which words (and thus POS) attention focuses on. Table 2 shows the top ten words located under peaks for both languages (and their corresponding frequency in the training corpus). In order to see if the attention mechanism does any better than learning corpus statistics, we need a baseline POS distribution for comparison. One possibility would be to simply compare the proportion of peaks under a given POS to the frequency of the same POS computed on tokens (as provided in Table 2). However, by doing so, we would assume that all tokens have the same length in the speech signal, which is not the case (verbs are longer than determiners for instance). Thus, for each spoken utterance of the test set, we sampled $50 * p$ random peak positions (p number of true detected peaks per utterance), and computed the POS distribution over such peaks (see 3a). We consider this as our baseline corpus distribution if attention peaks were to occur randomly.

sal VERB tag, thus the high proportion of verbs in the Japanese dataset.

[7]Uses the first order difference of the input array - see https://github.com/lucashn/peakutils.



(a)

(b)

Fig. 3: (a) Baseline POS distribution if attention peaks were to occur randomly. (b) POS distribution of words under detected attention peaks. English (blue) and Japanese (red).

4.1.1. *English*

We notice (Fig. 3b) that the attention mechanism of the English model primarily focuses on NOUNS: 82% of the peaks are located above nouns. This is far above corpus frequency, which is 47%. The attention mechanism considers neither determiners (DET) nor adpositions (ADP) nor adjectives (ADJ) as relevant as only 0.6%, 3%, and 2.85% are highlighted, where corpus frequencies would predict 7%, 8%, and 8% respectively. Verbs (VERB) are half as often highlighted as corpus frequency would predict, meaning attention barely relies on such words to make its prediction.

4.1.2. *Japanese*

The Japanese attention mechanism clearly makes use of particles[8] (PRT): 45.77% of the peaks are located above such words where corpus frequency would predict 16.9%. In fact, 6 of the top ten words are particles (see Table 2). Moreover, 17.83% of the peak highlight speech segments corresponding to the GA particle, well before nouns: GA is a particle that is used to indicate that the preceding word is the subject of the sentence. Thus, detecting such a particle is most useful, as the preceding word surely is the main object of the target image. The Japanese attention mechanism also seems to rely on nouns as 47.79% of peaks are located above nouns. One could argue this value is not very different from corpus frequency: 47.42%. However, if such POS were to hinder prediction, we would expect the attention mechanism to lower the number of peaks above such words, such as the model did for verbs or adjectives, which is not the case here, meaning NOUNS are useful for the model's prediction.

4.1.3. *Child language acquisition and noun-bias*

When learning their native language, it has been theorised that children exhibit a noun-bias [23]:[9] that is, in most languages children learn nouns before any other catagory. We notice that both models exhibit such language-general behaviour and favour nouns over other categories. Also, we showed that our Japanese model develops

[8]Particles are small suffixed grammatical words.

[9][23] states that "words that refer to concepts are easy to learn because the child has already formed object concepts, and need only match words and concepts".

a language-specific behaviour when mainly focusing on GA particles. [24] demonstrated that Japanese toddlers also make use of GA to segment speech before any other particle. The *noun-bias* phenomenon in our corpus can be explained by two factors: first, images in our corpus display many objects, thus prompting annotator to use more nouns than verbs; second, VGG vectors (used to encode images) are only trained to detect objects and not actions.

4.2. Attention above word beginnings or word endings?

	Beginning	Middle-Beg.	Middle-End	End
EN	6.19	9.14	39.24	**45.42**
JA	27.90	18.70	17.58	**35.80**

Table 3: Position of attention peaks above words for English (EN) and Japanese (JA).

We analysed above which part of words peaks are located. We divided each word beneath a peak into 4 equal parts and counted the percentage of peaks located above a given category (see Table 3). We notice that peaks in our English model are mainly located on the second half of the words. This phenomenon is coherent with Slobin's [25] Operating Principles favoring language acquisition stating that children *"pay attention to the ends of words"*. Peaks in Japanese are located at word endings but also at word beginnings. It seems the very beginning of some particles is able to trigger an attention peak.

5. IMAGES AS PIVOTS FOR CROSS-LINGUAL SPEECH RETRIEVAL?

We have seen in previous section that attention focuses on nouns and Table 2 suggests that these nouns correspond to the main concept of the paired image. To confirm this trend, we experiment on a cross-lingual speech-to-speech retrieval task using images as pivots.

This possibility was introduced in [8], but required training jointly or alternatively two speech encoders within the same architecture and a parallel bilingual speech dataset while we experiment with separately trained models for both languages. In [8], a parallel corpus was needed as the loss functions adopted try to minimise either the distance between captions in two languages or the distance between captions in two languages and the associated image as pivot. As our approach uses two monolingual models, we do not need a parallel corpus. Each monolingual model can be trained on its own dataset featuring images and their spoken description. The approach is the following: we first select a set of pivot images never seen by any of the monolingual models before. We encode these images with the image encoder of each language.[10] Then, for each speech utterance query in a source language u_{src} (English for instance), we find the nearest speech utterance in the target language u_{tgt} (Japanese for instance) which minimises the cumulated distance $d(u_{src}, i) + d(i, u_{tgt})$ among all pivot images i.

To make sure no parallel dataset is used, we trained a new English model on the first half of the train set, and a new Japanese model on the second half. We evaluated our approach on 1k captions of our test corpus to be comparable with [8].[11] At the time of the evaluation, 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 and give its median rank $\widetilde{r}$. We report our results in Table 4 as well as results from [8] who performed speech-to-speech retrieval using crowd-sourced spoken captions in English and Hindi.

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

Table 4: Results on English (EN) to Japanese (JP) and Japanese to English speech-to-speech retrieval (subset of 1k captions). For comparison, we report [8]'s results on English to Hindi (HI) and Hindi to English speech-to-speech retrieval. Chance scores are R@1=.001, R@5=.005, and R@10=.01. Chance for median rank $\widetilde{r}$ is 500.5.

Our results are surprisingly high given the fact we did not train a bilingual model but used the output of two monolingual models never trained to solve such a task. Nevertheless, it is also important to mention that [8] experimented on real speech with multiple speakers while we used synthetic speech with only one voice. Table 5 shows an example of top-1 retrieved Japanese sentences for 2 English queries.

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

Table 5: Example of semantically related captions. English (EN) query and retrieved Japanese caption (JA) and its translation (TRANS).

6. CONCLUSION

In this paper we showed that attention in a neural model of visually grounded speech mainly focuses on nouns. We also showed that this behaviour holds true for two very typologically different languages such as English and Japanese and that attention could also develop language-specifc mechanisms to detect relevant information in one of the languages. We also provided evidence that it is possible to perform speech-to-speech retrieval with images as pivots using the output of two independently trained monolingual models. In future work, we would like to validate our methodology on a bilingual dataset featuring real voices and try to extract a bilingual speech-to-speech dictionary using attention peaks as anchor points.

Ultimatly, we would like to emphasise the paramount importance of using other languages than English when trying to analyse the linguistic representations learnt by neural networks so as to understand if the models encode language specific or language general information, and thus better understand their strengths and weaknesses.

7. ACKNOWLEDGEMENTS

We thank G. Chrupała and his team for sharing their code and dataset, as well as for helping us with technical issues.

[10] Since both image encoders (from English and Japanese) are trained separately, they do not lead to the same representation of an image.

[11] We did not perform evaluation on the full $25000_{EN} \times 25000_{JP}$ distance matrices where each source query is associated with 5 target captions. Instead, we randomly sub-sampled ten $1000_{EN} \times 1000_{JP}$ distance matrices so that there would be only one target caption for each query in order to compare our results with [8]. Results are averaged over 10 random samples.

8. REFERENCES

[1] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., Cham, 2014, pp. 740–755, Springer International Publishing.

[3] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Scottsdale, Arizona, USA, December 2015, pp. 237–244.

[4] 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)*. 2017, pp. 613–622, Association for Computational Linguistics.

[5] William Havard, Laurent Besacier, and Olivier Rosec, "Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set," in *Proc. GLU 2017 International Workshop on Grounding Language Understanding*, 2017, pp. 42–46.

[6] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi, "Stair captions: Constructing a large-scale japanese image caption dataset," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2017, pp. 417–421, Association for Computational Linguistics.

[7] Afra Alishahi, Marie Barking, and Grzegorz Chrupała, "Encoding of phonology in a recurrent neural model of grounded speech," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. 2017, pp. 368–378, Association for Computational Linguistics.

[8] David Harwath, Galen Chuang, and James R. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 4969–4973.

[9] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, Eds., Cham, 2018, pp. 659–677, Springer International Publishing.

[10] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata, "Image pivoting for learning multilingual multimodal representations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 2839–2845, Association for Computational Linguistics.

[11] Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi, "Representation of linguistic form and function in recurrent neural networks," *Comput. Linguist.*, vol. 43, no. 4, pp. 761–780, Dec. 2017.

[12] David Harwath and James Glass, "Learning word-like units from joint audio-visual analysis," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 506–517, Association for Computational Linguistics.

[13] Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," in *INTERSPEECH*, 2017.

[14] David F. Harwath, Antonio Torralba, and James R. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 1858–1866.

[15] Emmanuel Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," *Cognition*, vol. 173, pp. 43 – 59, 2018.

[16] Eleanor Jack Gibson, *Principles of perceptual learning and development*, The century psychology series. Prentice-Hall, 1969.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of ICLR 2015*, 2015, pp. 1–14.

[18] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *INTERSPEECH*, 2017.

[19] Thomas Kisler, Uwe Reichel, and Florian Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326 – 347, 2017.

[20] Helmut Schmid, "Probabilistic part-of-speech tagging using decision trees," in *New Methods in Language Processing*, Daniel Jones and Harold Somers, Eds., Studies in Computational Linguistics, pp. 154–164. UCL Press, London, GB, 1997.

[21] Graham Neubig, Yosuke Nakata, and Shinsuke Mori, "Pointwise prediction for robust, adaptable japanese morphological analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011, pp. 529–533, Association for Computational Linguistics.

[22] Slav Petrov, Dipanjan Das, and Ryan McDonald, "A universal part-of-speech tagset," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, Eds., Istanbul, Turkey, may 2012, European Language Resources Association (ELRA).

[23] Dedre Gentner, "Why nouns are learned before verbs: Linguistic relativity versus natural partitioning," *Language*, vol. 2, pp. 301–334, 1982.

[24] Etsuko Haryu and Sachiyo Kajikawa, "Use of bound morphemes (noun particles) in word segmentation by japanese-learning infants," *Journal of Memory and Language*, vol. 88, no. C, pp. 18–27, 2016.

[25] C.A. Ferguson and D.I. Slobin, *Studies of child language development*, New York : Holt, Rinehart and Winston, 1973.