Modelling Human Word Learning and Recognition Using Visually Grounded Speech

Merkx, Danny; Scholten, Sebastiaan; Frank, Stefan L.; Ernestus, Mirjam; Scharenborg, Odette

DOI
10.1007/s12559-022-10059-7

Publication date
2022

Document Version
Final published version

Published in
Cognitive Computation

Cognitive Computation
https://doi.org/10.1007/s12559-022-10059-7

Modelling Human Word Learning and Recognition Using Visually Grounded Speech

Danny Merkx[1,2,3] · Sebastiaan Scholten[4] · Stefan L. Frank[1] · Mirjam Ernestus[1] · Odette Scharenborg[4]

Received: 6 January 2022 / Accepted: 25 September 2022

Abstract
Many computational models of speech recognition assume that the set of target words is already given. This implies that these models learn to recognise speech in a biologically unrealistic manner, i.e. with prior lexical knowledge and explicit supervision. In contrast, visually grounded speech models learn to recognise speech without prior lexical knowledge by exploiting statistical dependencies between spoken and visual input. While it has previously been shown that visually grounded speech models learn to recognise the presence of words in the input, we explicitly investigate such a model as a model of human speech recognition. We investigate the time course of noun and verb recognition as simulated by the model using a gating paradigm to test whether its recognition is affected by well-known word competition effects in human speech processing. We furthermore investigate whether vector quantisation, a technique for discrete representation learning, aids the model in the discovery and recognition of words. Our experiments show that the model is able to recognise nouns in isolation and even learns to properly differentiate between plural and singular nouns. We also find that recognition is influenced by word competition from the word-initial cohort and neighbourhood density, mirroring word competition effects in human speech comprehension. Lastly, we find no evidence that vector quantisation is helpful in discovering and recognising words, though our gating experiment does show that the LSTM-VQ model is able to recognise the target words earlier.

Keywords Computational modelling · Human speech recognition · Multi-modal learning · Deep learning · Vector quantisation

Introduction

Infants initially have little understanding of what is being said around them, and yet at approximately 9 months old are able to produce their first words. When they start producing their first multi-word utterances around 18 months, they can already produce about 45 words and comprehend many more [1, 2]. One of the challenges infants face is that speech does not contain neat breaks between words, which would allow them to segment the utterance into words. To complicate things further, words might be embedded in longer words (e.g. *ham* in *hamster*) and furthermore, no two realisations of the same spoken word are ever the same due to speaker differences, accents, co-articulation and speaking rate, etc. [3]. In this study, we investigate whether a computational model of speech recognition inspired by infant learning processes can learn to recognise words without prior linguistic knowledge.

Cognitive science has long tried to explain our capacity for speech comprehension through computational models (see [4] for an overview). Models such as Trace [5], Cohort [6], Shortlist [7], Shortlist B [8] and FineTracker [9] attempt to explain how variable and continuous acoustic signals are mapped onto a discrete and limited-size mental lexicon. These models all assume that the speech signal is first mapped to a set of pre-lexical units (e.g. phones, articulatory features) and then to a set of lexical units (words). The exact set of units is predetermined by the model developer, avoiding the issue of learning what these units are in the

✉ Danny Merkx
d.g.m.merkx@tudelft.nl

1 Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

2 Department of Intelligent Systems, Delft University of Technology, Delft, The Netherlands

3 Dutch National Police, The Hague, The Netherlands

4 Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands

Published online: 15 October 2022

first place. Even the recently introduced DIANA model [10], which does away with fixed pre-lexical units, uses a set of predetermined lexical units.

While all these models have proven successful at explaining behavioural data from listening experiments, they all require prior lexical knowledge in the form of a fully specified set of (pre-)lexical units. In contrast, infants learn words without prior lexical knowledge (or, arguably, any other linguistic knowledge) as well as without explicit supervision. A viable computational model should simulate word learning in a similar manner.

We take inspiration from the way infants learn language in order to model human word learning and recognition in a more cognitively plausible and 'human-like' manner. While learning language, children are exposed to a wide range of sensory experiences beyond purely linguistic input. On the other hand, current computational models of word learning and recognition are often limited to linguistic input. Using a multi-modal model, we aim to show that it is possible learn to recognise words without prior lexical knowledge and explicit supervision if the model is exposed to sensory experiences beyond speech. While there are many sensory experiences that could contribute to language learning, we focus on the most prominent of the human senses: vision. The model that we investigate in the current work exploits visual context in order to learn to recognise words in speech without supervision or prior lexical knowledge.

Visually Grounded Speech

Humans have access to multiple streams of sensory information besides the speech signal, perhaps most prominently the visual stream. It has been suggested that infants learn to extract words from speech by repeatedly hearing words while seeing the associated objects or actions [11], and indeed speech is often used to refer to and describe the world around us. For instance, parents might say 'the ball is on the table' and 'there's a ball on the floor' etc., while consistently pointing towards a ball.

Visually Grounded Speech (VGS) models are speech recognition models inspired by this learning process. The basic idea behind VGS models (e.g. [12–14]) is to make use of co-occurrences between the visual and auditory streams. For instance, from the sentences 'a dog playing with a stick' and 'a dog running through a field' along with images of these scenes, a model could learn to link the auditory signal for 'dog' to the visual representation of a dog because they are common to both image-sentence pairs. This allows the model to discover words, that is, to learn which utterance constituents are meaningful linguistic units. While there is a wide variety of VGS models, they all share the common concept of combining visual and auditory information in a common multi-modal representational space in which the similarity between matching image-sentence pairs is maximised while the similarity between mismatched pairs is minimised.

The potential of visual input for modelling the learning of linguistic units has long been recognised. In 1998, Roy and Pentland introduced their model of early word learning [15]. While many models at the time (and even today) relied on phonetic transcripts or written words, they implemented a model that learns solely from co-occurrences between the visual and auditory inputs. This model builds an 'audio-visual lexicon' by finding clusters in the visual input and looking for reoccurring units in the acoustic signal. It performs many tasks that are still the focus of research today: unsupervised discovery of linguistic units, retrieval of relevant images, and generation of relevant utterances. However, the model was limited to colours and shapes (utterances such as 'this is a blue ball') and has not been shown to learn from more natural, less restricted input.

The tasks performed by Roy and Pentland's model involve challenges for both computer vision and natural language processing. Advances in both fields have led to renewed interest in multi-modal learning, and with it increased the need for multi-modal datasets. In 2013, Hodosh, Young and Hockenmaier introduced Flickr8k [16], a database of images accompanied by written captions describing their contents, which was quickly followed by similar databases such as MSCOCO Captions [17]. These datasets are now widely used for image-caption retrieval models (e.g. [18–24]) and caption generation (e.g. [19, 25]).

Harwath and Glass collected spoken captions for the Flickr8k database and used it to train the first neural network-based VGS model [26]. Since then, there have been many improvements to the model architecture ([27–33]), as well as new applications of VGS models such as semantic keyword spotting ([14, 34, 35]), image generation [36], recovering of masked speech [37], and even the combination of speech and video [38].

Many studies have since investigated the properties of the representations learned by such VGS models (e.g. [13, 39–42]). Perhaps the most prominent question is whether words are encoded in these utterance embeddings even though VGS models are not explicitly trained to encode words and are only exposed to complete sentences. The VGS model presented in [31] showed that representations of a speech unit and a visual patch are often most similar when the visual patch contains the speech unit's visual referent. In [28, 29], the authors show that VGS models encode the presence of individual words that can reliably be detected in the resulting sentence representation.

Räsänen and Khorrami [43] made a VGS model that was able to discover words from even more naturalistic input than image captions: recordings from head-mounted cameras worn by infants during child-parent interaction. The

authors showed that their model was able to learn utterance representations in which several words (e.g. 'doggy', 'ball') could reliably be detected. Even though their model used visual labels indicating the objects the infants were paying attention to rather than the actual video input, this study is an important step towards showing that VGS models can acquire linguistic units from actual child-directed speech.

While the presence of individual words is encoded in the representations of a VGS model, the model does not explicitly yield any segmentation or discrete linguistic units. A technique which allows for the unsupervised acquisition of such discrete units is Vector Quantisation (VQ). VQ layers were recently popularised by [44], who showed that these layers could efficiently learn a discrete latent representational space. Harwath, Hsu and Glass [13] have recently applied these layers in a VGS model, and showed that their model learned to encode phones and words in its VQ layers.

Havard and colleagues went beyond simply detecting the presence of words in sentence representations: they presented isolated nouns to a VGS model trained on whole utterances, and showed that the model was able to retrieve images of the nouns' visual referents [45]. This shows that their model does not merely encode the presence of these nouns in the sentence representations, but actually 'recognises' individual words and learns to map them onto their visual referents. So, regarding the example mentioned above, the model learned to link the auditory signal for 'dog' to the visual representation of a dog.

However, the model by Havard and colleagues [45] was trained on synthetic speech. Word recognition in natural speech is known to be more challenging, as shown for instance by a large performance gap between VGS models trained on synthetic and real speech [28]. Dealing with the variability of speech is an important aspect of human speech recognition. If VGS models are to be plausible as computational models of speech recognition, it is important that these models implicitly learn to extract words from natural speech.

Current Study

The goal of this study is to investigate whether a VGS model discovers and recognises words from natural, as opposed to synthetic, speech. We furthermore go beyond earlier work because we investigate the model's cognitive plausibility by testing whether its word recognition performance is affected by word competition known to take place during human speech comprehension. We aim to answer the following questions:

1. Does a VGS model trained on natural speech learn to recognise words, and does this generalise to isolated words?
2. Is the model's word recognition process affected by word competition?
3. Does the model learn the difference between singular and plural nouns?
4. Does the introduction of VQ layers for learning discrete linguistic units aid word recognition?

Our **first** experiment is a continuation of our previous work [46] and the work by Havard et al. [45]. As in [45], we present isolated target words to the VGS model and measure its word recognition performance by looking at the proportion of retrieved images containing the target word's visual referent. If the model is indeed able to recognise a word in isolation, it should be able to retrieve images depicting the word's visual referent, indicating that the model has learned a representation of the word from the multi-modal input. Whereas previous work focused on the recognition of nouns, we also include verbs as our target words.

For this experiment, we collect new speech data, consisting of words pronounced in isolation. On the one hand, such data can be thought of as 'cleaner' than words extracted from sentences (as in [46]) due to the absence of co-articulation. On the other hand, the model was trained on words in their sentence context, co-articulation included, and might have learned to rely on this contextual information too heavily to also recognise words in isolation. Thus, to answer our first research question, we investigate whether our VGS model learns to recognise words independently of their context. Furthermore, we investigate whether linguistic and acoustic factors affect the model's recognition performance similarly to human performance. For instance, we know that faster speaking negatively impacts human word recognition (e.g. [47]).

In our **second** experiment we investigate the time course of word recognition in our VGS model. This allows us to test whether the model's word recognition performance is affected by word competition as is known to take place during human speech comprehension. For this experiment, we look at two measures of word competition: word-initial cohort size and neighbourhood density. In the Cohort model of human speech recognition [6], the incoming speech signal is mapped onto phone representations. These activated phone representations activate every word in which they appear. As more speech information becomes available, activation reduces for words that no longer match the input. The word that best matches the speech input is recognised. The number of activated or competing words is called the word-initial cohort size and plays a role in human speech processing: the larger the cohort size (i.e. the more competitors there are), the longer it takes to recognise a word [48]. Words with a denser neighbourhood of similar-sounding words are also harder to recognise as they compete with more words [49].

We also use our model to test the interaction between neighbourhood density and word frequency. Several studies have investigated this interaction, with inconclusive results. In a gating study, Metsala [50] found an interaction where recognition was facilitated by a dense neighbourhood for low-frequency words and by a sparse neighbourhood for high-frequency words. Goh et al. [51] found that response latencies in word recognition were shorter for words with sparser neighbourhoods. They furthermore found a higher recognition accuracy for sparse-neighbourhood high-frequency words as opposed to the other conditions (i.e. sparse-low, dense-high, dense-low). This means that, unlike Metsala, they found no facilitatory effect of neighbourhood density for low-frequency words. Others found no interaction between lexical frequency and neighbourhood density at all [52, 53].

For this experiment, we use a gating paradigm, a well-known technique borrowed from human speech processing research (e.g. [54, 55]). In the gating experiment, a word is presented to the VGS model in speech segments of increasing duration, that is, with an increasing number of phones, and the model is asked to retrieve an image of the correct visual referent on the basis of the speech signal available so far. We then analyse the effects of word competition and several control factors on word recognition performance.

In our **third** experiment we investigate whether our VGS model learns to differentiate between singular and plural instances of nouns. By the same principle of co-occurrences between the visual and auditory streams that allows the model to discover and recognise nouns, it may also be able to differentiate between their singular and plural forms. We test this by presenting both forms of all nouns to the model, and analysing whether the retrieved images contain single or multiple visual referents of that noun.

Our **fourth** question investigates VQ, a technique that was recently first applied to VGS models by Harwath, Hsu and Glass [13]. Their model acquired discrete linguistic units, including words. However, it is still an unanswered question whether such VQ-induced word units also aid the recognition of words in isolation. If they do, the addition of VQ layers should improve word recognition results of our VGS model. Havard, Chevrot and Besacier [30] improved retrieval performance of their VGS model by providing explicit word boundary information, thereby showing that knowledge of the linguistic units is indeed beneficial to the model. Rather than explicitly providing word boundary information, VQ layers allow units to emerge in an end-to-end fashion. Because prior knowledge of word boundaries is not cognitively plausible, VQ layers are a more suitable approach for our cognitive model. To investigate if the introduction of VQ layers indeed aids word recognition, all our experiments compare the baseline VGS model to a VGS model with added VQ layers.

To foreshadow our results, we find that (1) our VGS model does learn to recognise words in isolation but performance is much higher on nouns than on verbs; (2) word recognition in the model is affected by competition similarly to humans; (3) the model can distinguish between singular and plural nouns to a limited extent; and (4) the use of VQ layers does not improve the model's recognition performance.

Methods

Visually Grounded Speech Model

Model Architecture

Our VGS model consists of two deep neural networks as depicted in Fig. 1; one to encode the images and one to encode the audio captions. The model is trained to embed both input streams in a common embedding space; its training goal is to minimise the cosine distance between image-caption pairs while maximising the distance between mismatched pairs. We do not fine-tune the hyper-parameters of the model but use the best parameters found in [18] — this is because it is not our current goal to improve the training task score but to perform experiments in order to learn more about the unsupervised discovery and recognition of words in a VGS model.

It is common practice to use a pre-trained image recognition network for the image branch of a VGS model (e.g. [13, 28, 35]). We use the ResNet-152 network [56], which is a pre-trained convolutional network that was trained on ImageNet [57], to extract image features. This is done by taking the activations of ResNets-152's penultimate fully connected layer by removing the final object-classification layer. Our image branch then is a single linear layer of size 2048 applied to these image features. Finally, we normalise the results to have unit L2 norm. The goal of the linear projection is to map the image features to the same 2048-dimensional embedding space as the audio representations. The image embedding $\mathbf{i}$ is given by:

$$\mathbf{i} = \frac{\mathbf{img}A^T + \mathbf{b}}{||\mathbf{img}A^T + \mathbf{b}||_2}, \tag{1}$$

where A and $\mathbf{b}$ are learned weight and bias terms, and $\mathbf{img}$ is the vector of ResNet-152 image features.

The audio branch consists of a 1-d convolutional neural network of size 6, stride 2 and 64 output channels, which

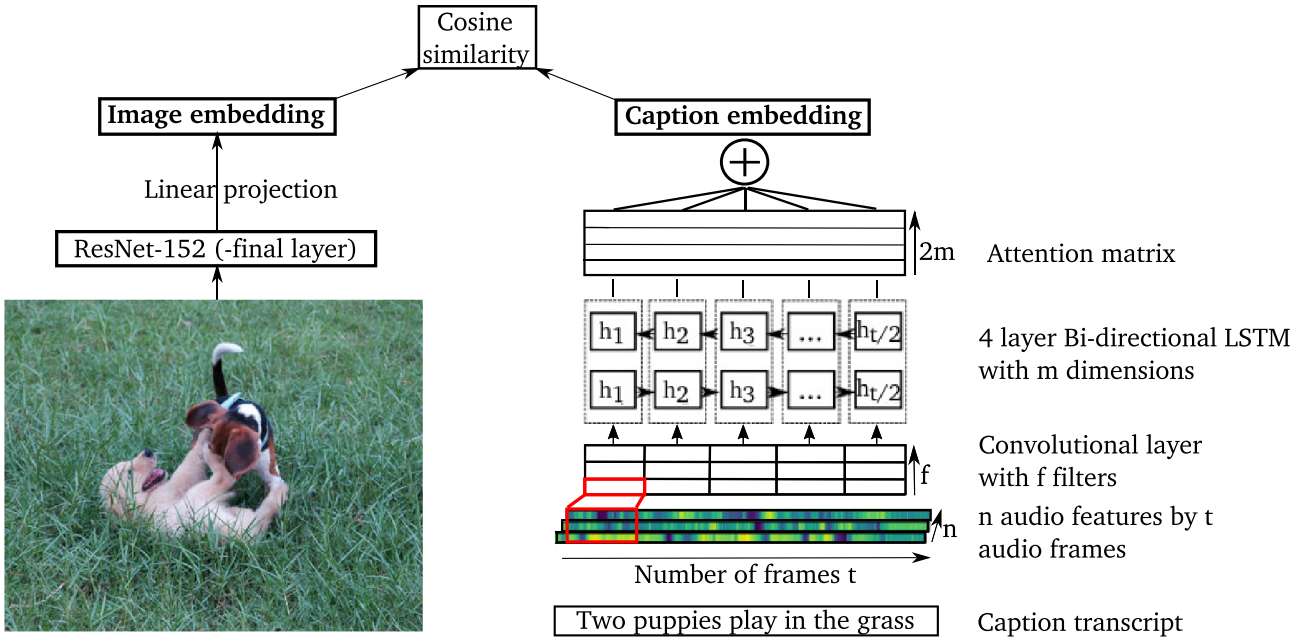

Fig. 1 Model architecture: the model consists of two branches with the image encoder depicted on the left and the caption encoder on the right. The audio features consist of 13 MFCC with 1st and 2nd order derivatives by t frames. Each LSTM hidden state $\mathbf{h}_t$ has 1024 features which are concatenated for the forward and backward LSTM into 2048-dimensional hidden states. Vectorial attention weighs and sums the hidden states resulting in the caption embedding. The linear projection in the image branch maps the image features to the same 2048-dimensional space as the caption embedding. We calculate the cosine similarity between the image and caption embedding

sub-samples the signal along the temporal dimension. The resulting features are fed into a 4-layer bi-directional Long Short Term Memory (LSTM) with 1024 units.[1] The 1024 bi-directional units are concatenated to create a 2048 feature vector. The self-attention layer computes a weighted sum over all the hidden LSTM states:

$$\mathbf{a}_t = \text{softmax}\,(V \tanh(W\mathbf{h}_t + \mathbf{b}_w) + \mathbf{b}_v), \tag{2}$$

where $\mathbf{a}_t$ is the attention vector for hidden state $\mathbf{h}_t$, and W, V, $\mathbf{b}_w$, and $\mathbf{b}_v$ indicate the weights and biases. The learnable weights and biases are implemented as fully connected linear layers with output sizes 128 and 2048, respectively. The applied attention is then the sum over the Hadamard product between all hidden states $(\mathbf{h}_1, ..., \mathbf{h}_t)$ and their attention vector:

$$\text{Att}\,(\mathbf{h}_1, ..., \mathbf{h}_t) = \sum_t \mathbf{a}_t \circ \mathbf{h}_t. \tag{3}$$

The resulting embeddings are normalised to have unit L2 norm. The caption embedding $\mathbf{c}$ is thus given by:

$$\mathbf{c} = \frac{\text{Att}\,(\text{LSTM}\,(\text{CNN}\,(\mathbf{a}_1, ..., \mathbf{a}_t)))}{||\,\text{Att}\,(\text{LSTM}\,(\text{CNN}\,(\mathbf{a}_1, ..., \mathbf{a}_t)))||_2}, \tag{4}$$

where $\mathbf{a}_1, ..., \mathbf{a}_t$ indicates the caption represented as t frames of MFCC vectors and Att, LSTM and CNN are the attention layer, stacked LSTM layers, and convolutional layer, respectively.

Next, we also implement a VGS model with added VQ layers [44]. We will refer to our regular model and the model with VQ layers as LSTM and LSTM-VQ models, respectively. Our implementation most closely follows [13], who were the first to apply these layers in a VGS model, and showed that their model learned discrete linguistic units. VQ layers consist of a 'codebook' which is a set of n-dimensional embeddings. A VQ layer discretises incoming input by mapping it to the closest embedding in the codebook and passing this embedding to the next layer:

$$VQ(\mathbf{x}) = \mathbf{e}_k, \text{ where } k = \text{argmin}_j ||\mathbf{x} - \mathbf{e}_j||_2, \tag{5}$$

where $\mathbf{x}$ is the VQ layer input and $\mathbf{e}_j$ are the codebook embeddings.

For the LSTM-VQ model we insert VQ layers in the LSTM stack after the first and after the second LSTM layer, with 128 and 2048 codes, respectively. We use two layers because in [13] this made a hierarchy of linguistic units emerge: The first layer best captured phonetic identity while

[1] In [29] we used a 3-layer Gated Recurrent Unit, but it has since then become practically feasible to train larger models on our hardware.

in the second layer, several codes emerged that were sensitive to specific words.

We use our own PyTorch implementation of the models and the VQ layer described here, adapted from our previous work presented in [18, 29], which is in turn most closely related to, and based on, the VGS models presented in [27, 28]. Our implementation and data can be found on https://github.com/DannyMerkx/speech2image/tree/CogComp2022.

Training Data

We train the model on Flickr8k [16], a well-known dataset of 8000 images from the online photo sharing platform Flickr.com, with five written English captions per image. Annotators were asked to 'write sentences that describe the depicted scenes, situations, events and entities (people, animals, other objects)' [16]. We use the spoken captions Harwath and Glass [26] collected by having Amazon Mechanical Turk (AMT) workers pronounce the original written captions. We use the data split provided by [19], with 6000 images for training and a development and test set of 1000 images each.

Image features are extracted by resizing all images while maintaining the aspect ratio such that the smallest side is 256 pixels. Ten crops of 224 by 224 pixels are taken, one from each of the corners, one from the middle and similarly for the mirrored image. We use ResNet-152 [56] to extract visual features from these ten crops and then average the features of the ten crops into a single vector with 2048 features.

The audio input consists of Mel Frequency Cepstral Coefficients (MFCCs). We compute the MFCCs using 25 ms analysis windows with a 10 ms shift. The MFCCs were created using 40 Mel-spaced filterbanks. We use 12 MFCCs and the log energy feature, and add the first and second derivatives resulting in 39-dimensional feature vectors. Lastly, we apply per-utterance cepstral mean and variance normalisation.

Training

The model is trained to embed the images and captions such that the cosine similarity between image and caption embeddings is larger for matching pairs than the similarity between mismatching pairs. The batch hinge loss L as a function of the network parameters θ is given by:

$$L(\theta) = \sum_{(\mathbf{c},\mathbf{i}),(\mathbf{c}',\mathbf{i}')\in B} \Big(\max(0, \cos(\mathbf{c},\mathbf{i}') - \cos(\mathbf{c},\mathbf{i}) + \alpha) + \max(0, \cos(\mathbf{i},\mathbf{c}') - \cos(\mathbf{i},\mathbf{c}) + \alpha) \Big), \tag{6}$$

where $(\mathbf{c},\mathbf{i}) \neq (\mathbf{c}',\mathbf{i}')$, B is a minibatch of matching caption-image pairs $(\mathbf{c},\mathbf{i})$, and the other caption-image pairs $(\mathbf{c}',\mathbf{i}')$ in the batch serve to create mismatching pairs: $(\mathbf{c},\mathbf{i}')$ and $(\mathbf{c}',\mathbf{i})$. We take the cosine similarity and subtract the similarity of the mismatching pairs from the matching pairs such that the loss is only zero when the matching pair is more similar than the mismatching pairs by a margin α, which was set to 0.2.

Training task performance is evaluated by caption-to-image and image-to-caption retrieval score Recall@N on the 1000-image test set. For these retrieval tasks, the caption embeddings are ranked by cosine distance to the image and vice versa, and Recall@N is the percentage of test items for which the correct image or caption was in the top N results. Furthermore, we evaluate the median rank of the correct image or caption.

Because the VQ operation is indifferentiable, a trick called *straight through estimation* is required to pass a learning signal to layers before the VQ layer [58]. Put simply, as there is no gradient for the VQ operation, the gradients for the VQ output are copied and used as an approximation of the gradients for the VQ input.

The VQ layer learns to make the codebook codes more similar to their inputs and vice versa. The first is accomplished by an exponential moving average. When a code is activated, it gets multiplied by a decay factor γ and summed with $(1-\gamma)\mathbf{x}$, where $\mathbf{x}$ is the input that activated the code. Making the inputs more similar to the codes is accomplished by a separate VQ loss, which is the mean squared error between each input and its closest code.

The networks are trained using Adam [59] with a cyclic learning rate schedule based on [55]. The learning rate schedule varies the learning rate smoothly between a minimum of 10^{-6} and maximum of 2×10^{-4}.

We train the regular LSTM-based network for 16 epochs. Following [13], we *warm start* the LSTM-VQ model by taking the trained LSTM network, inserting the VQ layers and training for another 16 epochs. While, unlike [13], we did not encounter a large performance loss for *cold started* networks, we did find that a cold started VQ network frequently suffered from codebook collapse [60]. This is an issue where suddenly all VQ inputs are mapped to only a few (often even just one) codes and from which the model never recovers.

We trained 20 VGS models of each type (with and without VQ) using different seeds for the pseudo-random number generator, to average over random effects of weight initialisation and training data presentation order.

Data Collection

Target Words

Word learning by visually grounded speech models exploits the fact that words in the speech signal tend to co-occur with visual referents in the corresponding images. We can therefore expect that any words the system learns to recognise

Table 1 Selected target nouns and verbs in order of occurrence in the training set. An * indicates nouns for which only the singular or plural form was recorded, + indicates words that were not included in the analysis because there were not enough images depicting their visual referent in the test set

Nouns		Verbs	
dog	man	play	run
boy	girl	jump	sit
woman	water*	hold	walk
shirt	ball	ride	climb
grass*	beach	smile	pose
snow*	group	catch	carry
street	rock	leap	perform
camera	bike	fly	dance
mountain	hat	swim	eat
pool	player	pull	hang
jacket	ocean	chase	slide
basketball	sand*	splash	point
car	building	kick	throw
soccer*	swing	fight	swing
football	sunglasses*	lie	lay
shorts*	park	laugh	ski
dress	table	surf	drive
hand	tree	fall	follow
lake	hill	race	roll
toy	baby	hit	reach
tennis*+	river	wade	lean
wave	snowboarder	push	bite
bench	game	spray	paddle
surfer	stick	light+	bend
team	skateboard	cross	raise

will be words with visual referents in the images. Hence, we limit our analysis to the recognition of nouns and verbs. We only look at high-frequency words that the model has had ample opportunity to learn to recognise.

We selected the 50 nouns and 50 verbs with the most frequent lemma in the Flickr8k database, excluding some words like 'air' and 'stand' as their referents appear in nearly every picture and, consequently, whether the words are recognised cannot be established. Other examples of rejected words are verbs such as 'try' for which it is not possible to set objective standards for the visual referent. The selected words are shown in Table 1.

To test word recognition performance, we present the selected target verbs and nouns in isolation. Two North American native speakers of English (one male, one female), not present in the Flickr8k database, were asked to read the target words out loud from paper. The words were recorded in isolation by asking the speakers to leave at least a second of silence in between words. To keep conditions close to those of the Flickr8k spoken captions (and other captioning databases collected through AMT), the speakers recorded the words at home using their own hardware. They were asked to find a quiet setting and record the words in a single session. They received a $20 gift card for their participation.

The nouns were presented in both their singular and plural form (where applicable)[2]. All verbs were recorded in root form, third person singular form, and progressive participle form. We did not record past tense forms as these are rarely, if ever, used in the image descriptions.

The speech data were recorded in stereo at 44.1kHz in Audacity. We down-sampled the utterances to 16kHz and converted them to mono to match the conditions of the Flickr8k captions, after which we applied the same MFCC processing pipeline used for the Flickr8k training data.

Image Annotations

We test whether the VGS model learned to recognise the recorded target words by presenting them to the model and checking whether the retrieved images contain the words' visual referents. The problem with this approach, however, is that Flickr8k contains no ground truth image annotations for such a test. The captions can serve as an indication: if annotators mention an action or object in the caption we can be reasonably sure it is visible in the picture. In contrast, it is definitely not the case that if an object or action is not mentioned, it is not in the picture. Hence, using captions as ground truth would lead to an underestimation of model performance.

We created a ground truth labelling for the visual referents of our target words by manually annotating the 1000 images in the Flickr8k test set for visual presence of each target word. For the nouns, we also indicate whether the visual referent occurred only once or multiple times in the images, allowing us to test whether the model learns to differentiate between plural and singular nouns.

There were two annotators, one covering the nouns and one the verbs. To check the quality of the annotations, the first author annotated a sample of 5% of the images. The inter-annotator agreement based on this sample was $\kappa = 0.70$ for verbs and $\kappa = 0.76$ for nouns.

Word Recognition

We take the retrieval of images containing a target word's visual referent as indicative of successful word recognition. As this is a retrieval task where multiple correct images can be found per word, we use precision@10 (P@10) to measure

[2] 'Shorts' and 'sunglasses' are syntactically plural, but we group them under the singular nouns as their use in the data is most often in reference to a single object.

word recognition performance, following [45]. That is, for each target word embedding we calculate the cosine similarity to all test image embeddings and retrieve the ten most similar images. P@10 is then the percentage of those images that contains the visual referent according to our annotations. We excluded two target words from this analysis as there were fewer than ten test images containing their visual referent. Although we annotated whether an image contains a single or multiple visual referents, unless stated otherwise, multiple visual referents were counted as correct for a singular noun and vice versa for the purpose of calculating P@10.

We also compute P@10 scores for two baseline models. Our *random* baseline is simply the averaged score over five randomly initialised and untrained VGS models. This results in a random selection of images but since some words' visual referents occur in dozens to hundreds of test images, the recognition scores are far from zero. Our *naive* baseline is the recognition score of a model that always retrieves the ten images with the highest number of visual referents (i.e. always the same ten images, selected separately for the nouns and verbs). Note that this baseline is not realistic and requires knowledge of the contents of the test set (namely the number of visual referents per image). Still, it is useful to compare our model performance to a model that has only a single response regardless of the input.

We then examine the influence of linguistic and acoustic factors on the model's word recognition performance as measured by P@10, using a Generalised Linear Mixed Model (GLMM) with beta-binomial distribution[3] and canonical logit link function. We used the *glmmTMB* package in R [61].

The GLMM examines the effects of signal duration (i.e. number of speech frames), speaking rate (number of phones per second), number of vowels, number of consonants, morphology (singular or plural)[4] and VQ (LSTM or LSTM-VQ model), with the VGS model's word recognition performance (P@10) as the outcome variable. As control variables, we furthermore include the (log-transformed) counts of the target word and its lemma in the training set as we expect better recognition for words that are seen more often during training. The correlation between lemma count and word count is .48, so they are expected to explain unique portions of variance. We also include speaker-ID to account for differences in recognition performance between the two speakers. Numbers of vowels and consonants are centred; all other non-categorical variables are standardised. VQ (LSTM $= -1$, LSTM-VQ $= 1$), morphology (plural $= -1$, singular $= 1$) and speaker ID (#1 $= -1$, #2 $= 1$) were sum coded.

The GLMM includes by-lemma and by-model (each of the 20 random initialisations) random intercepts. We first included all fixed effects that vary within lemma or model-ID as by-lemma or by-model random slopes but this model was unable to converge. As a maximal model is thus not possible, we reduced the model until it converged: We tried a zero-correlation-parameter GLMM, which also did not converge. Next, we split the GLMM into one with only the by-lemma and one with only the by-model random slopes (uncorrelated). The by-model GLMM resulted in a singular fit for the speaker ID, morphology, and VQ random slopes. After removing these by-model slopes, the combined GLMM, with all remaining uncorrelated by-lemma and by-model slopes, converged. None of the removed random slopes could be added back into the combined GLMM without causing convergence issues. The final GLMM formula is:

```
p@10 ~ speaking rate + duration + lemma count + word count + #vowels + #consonants + VQ + speaker id + morphology + (1 + speaking rate + duration + word count + #vowels + #consonants + VQ + speaker id + morphology || lemma) + (1 + speaking rate + duration + lemma count + word count + #vowels + #consonants || model id)
```

, where the double pipe symbol (`||`) means that correlations between random slopes are not estimated.

Word Competition

We perform a gating experiment to investigate word competition in our models. We present the models with the target words in segments of increasing length, using one gate per phone. Simply put, if the target word is 'dog' with the phones /d-ɔ-g/, we evaluate performance after the model has processed /d/, /d-ɔ/, and finally the whole word /d-ɔ-g/. Performance is measured in P@10 as described in '2.3'.

For the gating experiment we need to know when each phone starts and ends. We use the Kaldi toolkit to make a forced alignment of our target words and their phonetic transcripts [62], taken from the CMU Pronouncing Dictionary available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

We define the word-initial cohort of a target word at a certain gate to be the set of words in the Flickr8k dataset that share the target's word-initial phone sequence up to the gate. That is, the number of words in the word-initial cohort equals the number of words that cannot be distinguished from the target given the sequence so far, and thus the number of words competing for recognition.

[3] Our P@10 data, which is discrete and has a floor of 0 and a ceiling of 10, is not suited for standard linear modelling. Our response variable is best described as a series of Bernoulli trails with successes and failures in terms of correct and incorrect retrieval.

[4] As seen in '3.1', word recognition results on the verbs were overall a lot worse than for the nouns so we decided not to continue our analysis on the verbs.

We define neighbourhood density as the number of words in Flickr8k that differ by exactly one phone from the target word [63]. These words are expected to compete for recognition and so affect word recognition. Research shows that words with a dense neighbourhood are harder to recognise than those with a sparse neighbourhood [49].

For both the word-initial cohort and the neighbourhood density, we use phonetic transcripts from the CMU pronouncing dictionary, which contains the transcripts for a total of 6431 words in the Flickr8k captions.

We use a GLMM to test whether the neighbourhood density and word-initial cohort size affect word recognition in our model. Furthermore, we are interested in three interaction effects: as previously discussed, we test the interactions between neighbourhood density and the word and lemma counts. The third interaction is between VQ and the number of phones processed so far (gate number). The VGS model with VQ layers is forced to map its inputs to discrete units even as early as the first gate. As the second VQ layer has been shown to learn discrete word-like representations [13], we might expect that words are recognised earlier, as would be indicated by a smaller effect of gate number for the LSTM-VQ model.

The GLMM's fixed effects are the neighbourhood density, gate number, the size of the word-initial cohort, VQ, morphology, the number of vowels and the number of consonants. Again we also add the occurrence frequencies of the target word and its lemma in the training set and speaker-ID to account for expected effects of training data frequency and speaker differences. The number of vowels, number of consonants and gate number are centred; all other non-categorical variables are standardised.

The GLMM has by-lemma and by-model random intercepts. We started with maximal by-lemma and by-model random slopes but had to reduce the complexity due to convergence issues, using the same procedure as described before. However, after removing all random slopes that yielded singular fits in the GLMM with only by-model random effects, the combined model (with by-model and by-lemma random effects) still failed to converge. We proceeded to use the variance estimates of the separate GLMMs to remove the smallest variance components until the combined GLMM converged. This led to the removal of all by-model random slopes and the by-lemma slopes for number of vowels and word count. The final GLMM formula for analysis of the gating experiment is:

```
p@10 ~ (lemma count + word count) * density + VQ * gate + initial cohort size + speaker id + morphology + #vowels + #consonants + (1 + density + VQ + gate + initial cohort size + speaker id + morphology + #consonants || lemma) + (1 | model id)
```

Table 2 Image-caption retrieval results on the Flickr8k test set. R@N is the percentage of items for which the correct image or caption was retrieved in the top N (higher is better) with 95% confidence interval. Med r is the median rank of the correct image or caption (lower is better). We compare our VGS models to previously published results on Flickr8k. '-' means the score is not reported in the cited work

Model	Caption to Image			
	R@1	R@5	R@10	med r
[26]	-	-	17.9±1.1	-
[28]	5.5±0.6	16.3±1.0	25.3±1.2	48
[29]	8.4±0.8	25.7±1.2	37.6±1.3	21
[36]	10.1±0.8	28.8±1.3	40.7±1.4	-
LSTM	12.5±0.2	33.8±0.3	46.8±0.3	12
LSTM-VQ	12.9±0.2	34.5±0.3	47.3±0.3	12
Model	Image to Caption			
	R@1	R@5	R@10	med r
[26]	-	-	24.3±2.7	-
[29]	12.2±2.0	31.9±2.9	45.2±3.1	13
[36]	13.7±2.1	36.1±3.0	49.3±3.1	-
LSTM	18.5±0.5	42.4±0.7	55.8±0.7	8
LSTM-VQ	19.6±0.6	45.4±0.7	58.1±0.7	7

Results

All results presented here are averaged over the 20 random initialisations of the VGS model. We first evaluate how well the models perform on the training task and compare their performance to other VGS models. The scores in Table 2 show the result for the speech caption-to-image and image-to-caption retrieval tasks. This indicates how well the model learned to embed the speech and images in the common embedding space. As expected, the VQ layers are beneficial to the VGS model's training task performance [13].

Word Recognition

In the first experiment, we presented isolated words to the model. Table 3 shows the average P@10 scores. The singular nouns are recognised best with P@10 scores

Table 3 Word recognition results for each noun and verb type for the trained models, the random model, and the naive baseline. In parentheses are the recognition scores when only evaluating the subset of target words that also have plural forms

			Baseline	
Morphology	LSTM	LSTM-VQ	Random	Naive
singular noun	.519(.479)	.529(.485)	.137	.278
plural noun	.479	.449	.140	.267
root verb	.185	.193	.082	.188
third-person verb	.176	.164	.078	.188
participle verb	.246	.260	.083	.188

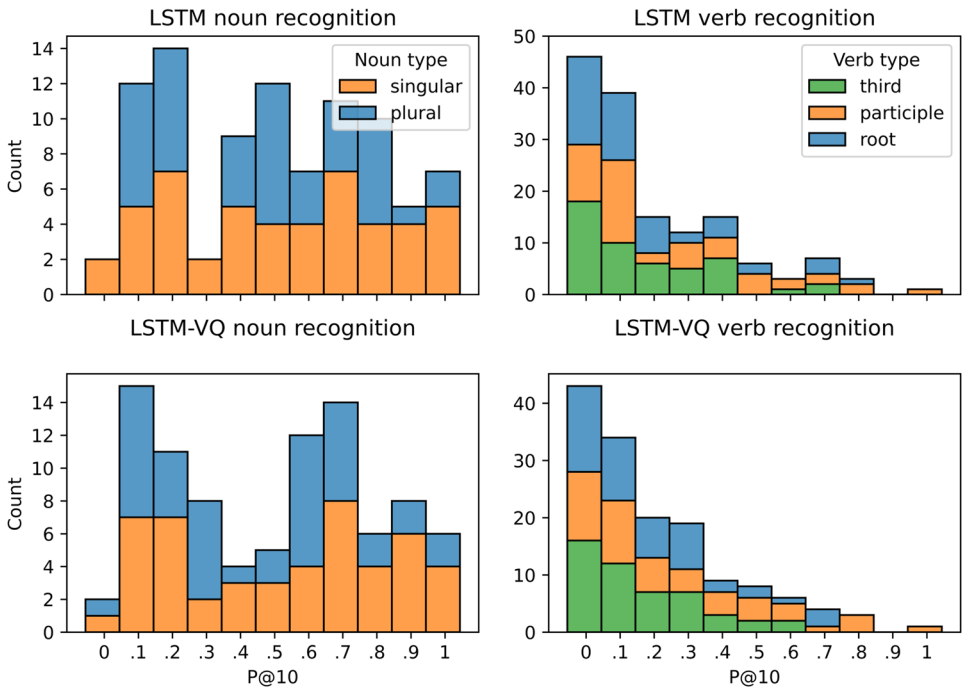

Fig. 2 Histograms of the word recognition experiment results for each word type

of .519 and .529 for the LSTM and LSTM-VQ model, respectively. This means that, on average, more than five out of the ten retrieved images contain the correct visual referent. For the plural nouns the average performance is .479 and .449 for the LSTM and LSTM-VQ model, respectively. However, seven target nouns have no plural form, so the scores for plural and singular nouns are not directly comparable. Therefore, we also calculate singular noun performance only on those words that also have a plural form. The results show that singular and plural forms are recognised equally well by the LSTM model. However, the LSTM-VQ model recognises plural target words slightly less accurately than singular words.

The histograms in Fig. 2 show the distribution of the P@10 scores by word type (noun or verb), morphology and whether the VGS model included VQ layers. This highlights that the recognition of the verbs is overall much worse than for the nouns: many verbs have a P@10 of zero, meaning they are not recognised at all. For the nouns on the other hand, only two words are not recognised at all. While both LSTM models outperform the random baseline on verb recognition, only on the participles is performance better than the naive baseline's, with scores over .7 on some words. As the recognition performance for the verbs is obviously a lot worse than for nouns, we continue our analysis on the nouns only.

Havard and colleagues [45] reported a median P@10 of 0.8 on 80 nouns (from the synthetic speech database MSCOCO), while our models achieve median P@10 scores of 0.6 and 0.5 on singular and plural nouns, respectively. Even though the models recognise most nouns and even their plural forms (with only two words per model not being recognised at all), this indicates a large drop in recognition performance going from the synthetic speech dataset in [45] to our natural speech. Note, however, that as Havard et al. used the most frequent nouns for their dataset (MSCOCO), the target words do not fully overlap with ours.

The results of the GLMM for the word recognition experiment are summarised in Table 4. Speaking rate and number of consonants have a significant effect on the VGS model's word recognition performance. The positive coefficient of the number of consonants indicates that words with more consonants are on average recognised better. The negative coefficient for speaking rate indicates that words are harder

Table 4 Estimated model effects for the word recognition GLMM and the results of Type III Wald χ^2 tests. Plural, LSTM and speaker 1 are the reference levels for Morphology, VQ and Speaker id respectively

Effect	Estimate	Std. error	χ^2	p
Intercept	−0.26	0.70	1.20	0.27
Speaking rate	−2.03	0.91	4.98	**0.03**
Duration	−0.88	0.60	2.14	0.14
Lemma count	1.98	0.70	7.97	**0.005**
Word count	0.33	0.40	0.69	0.41
#Vowels	1.33	1.35	0.98	0.32
#Consonants	2.06	0.81	6.46	**0.01**
VQ	0.02	0.04	0.34	0.56
Speaker id	−0.37	0.25	2.13	0.14
Morphology	−0.28	0.44	0.42	0.52

Table 5 Estimated model effects for our post-hoc testing of interaction effects and the results of Type III Wald χ^2 tests. LSTM and speaker 1 are the reference levels for VQ and Speaker id respectively. Plural and Participle are the Morphology reference levels for the noun and verb models respectively

Effect	Estimate	Std. error	χ^2	p
Nouns				
VQ	0.03	0.01	3.69	0.06
Word count:VQ	0.10	0.02	23.17	**<0.001**
Morphology				
Singular	1.34	0.86	2.45	0.12
Singular:VQ	0.12	0.02	38.42	**<0.001**
Verbs				
VQ	-0.04	0.01	11.02	**<0.001**
Word count:VQ	0.07	0.01	38.42	**<0.001**
Morphology				
Root	−0.05	0.22		
Third	0.46	0.33	6.85	**0.03**
Root:VQ	−0.07	0.02		
Third:VQ	−0.002	0.02	30.86	**<0.001**

to recognise if they are spoken faster. Unsurprisingly, lemma count also has a significant effect on word recognition: lemmas that were seen more often during training are recognised better. The results further confirm that plural and singular nouns are recognised equally well and that there is no difference in recognition performance between the two speakers.

While overall these results show no difference in word recognition performance between the LSTM-VQ and the LSTM models, it is notable that only LSTM-VQ has a performance difference between singular and plural nouns. Similarly, LSTM-VQ performs best on the participle verb form and worse on the third person and root forms. Third person and root verbs are less frequent than participles, and plural nouns are less frequent than singulars. Hence, it may be the case that the codebook simply learns to encode frequent words better, and struggles with the less frequent word(form)s.

To further investigate whether the VQ models indeed recognise frequent words more accurately, we performed a post hoc test where we refit the word recognition GLMM with an interaction between VQ and word count and between VQ and morphology. We fit separate GLMMs on the noun and verb targets, the results of which can be seen in Table 5. We find the expected interactions between VQ and morphology where recognition on the less frequent word forms (plural, third and root) is worse than on the more frequent forms (singular, participle) for the VQ network. Furthermore, we also find positive interactions between word count and VQ, further indicating that frequency of exposure has a greater effect on the LSTM-VQ models than on the LSTM models.

Table 6 Estimated model effects for the gating GLMM and the results of Type III Wald χ^2 tests. Plural, LSTM and speaker 1 are the reference levels for Morphology, VQ and Speaker id respectively

Effect	Estimate	Std. error	χ^2	p
Intercept	−0.71	0.24	9.10	**0.003**
Lemma count	0.87	0.20	18.1	<0.001
Word count	0.06	0.14	0.17	0.68
#Vowels	−0.08	0.29	0.07	0.79
#Consonants	0.57	0.21	7.42	**0.006**
Density	0.51	0.20	6.60	**0.01**
Gate	0.25	0.08	11.13	<0.001
Initial cohort	−0.98	0.20	23.0	<0.001
Morphology	−0.02	0.23	0.01	0.92
VQ	−0.09	0.05	3.18	0.07
Speaker id	0.21	0.14	2.36	0.12
Lemma count:density	0.19	0.13	2.09	0.15
Word count:density	−0.20	0.10	4.09	**0.04**
VQ:gate	0.03	0.01	11.61	<0.001

Word Competition

The results of the GLMM for the word competition experiment are summarised in Table 6. Of the fixed effects of interest, neighbourhood density, gate number, word-initial cohort size and number of consonants have significant effects on word recognition performance. Furthermore, we found significant interaction effects between word count and neighbourhood density, and between VQ and gate number.

As in the previous GLMM analysis, the number of consonants has a positive effect. The gate number (number of phones processed so far) also has a positive effect: unsurprisingly, the model is better able to recognise the target word as more of the word has been presented. This effect is modulated by the presence of VQ layers, where the negative coefficient indicates that the effect of gate is slightly smaller in the LSTM-VQ than in the LSTM models. There is a significant negative effect of word-initial cohort size. This means recognition performance is lower the more candidates there are. While neighbourhood density has an overall positive effect on word recognition, care should be taken in interpreting this effect in light of the negative interaction with word count. The positive effect would indicate that words with a higher neighbourhood density are recognised better; however, the interaction indicates this effect decreases with higher word count and might become negative for the most frequent words.

Plurality

Using the plurality annotations of the visual referents for the noun target words, we test whether the VGS models actually differentiate between singular and plural nouns. That

Table 7 Confusion matrices for singular and plural nouns indicating how many of the correctly retrieved images contained only one or multiple visual referents to the target word

Model	#refs in image	Noun morphology	
LSTM		singular	plural
	one	3048 (57%)	2940 (51%)
	multiple	2281 (43%)	2881 (49%)
LSTM-VQ		singular	plural
	one	2857 (56%)	2631 (49%)
	multiple	2278 (44%)	2754 (51%)

is, if we present it with a plural noun, does it return pictures with multiple visual referents? For this we first select only those target words which have both a plural and singular form. Then, we only keep those words which have at least ten images depicting a single visual referent and ten images with multiple visual referents. So, in theory the VGS models can achieve a perfect P@10 score on these words while also perfectly distinguishing between singular and plural nouns. This results in a final target word set of 28 nouns.

Table 7 shows the confusion matrices for the LSTM and LSTM-VQ models, with numbers of single- versus multiple-referent images returned when the model is presented with a singular versus plural target word. We see that both VGS models, when presented with singular nouns, more often return images with a single referent than with multiple referents. When presented with plural nouns, this difference decreases and, for LSTM-VQ, even reverses (LSTM: $\chi^2(1) = 49.8, p < 0.0001, N = 11,150$; LSTM-VQ: $\chi^2(1) = 48.1, p < 0.0001, N = 10,520$).

Recognition of plural nouns critically depends on the plural suffix, as this is what indicates whether a target

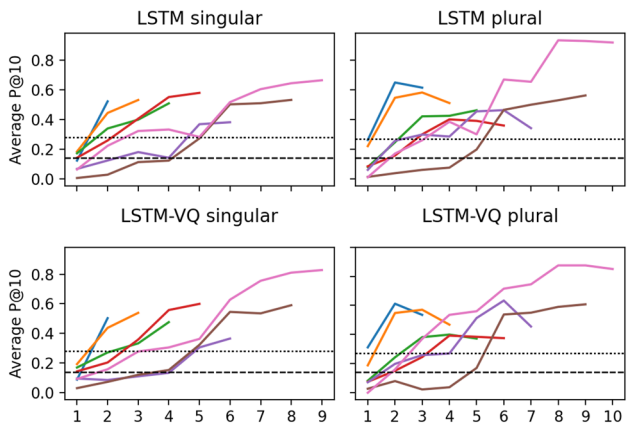

Fig. 3 Recognition scores as a function of the gate number (the number of phones processed so far). The solid lines represent averaged P@10 scores over words with an equal number of phones (the length and colour of each line indicate the number of phones). The dotted and dashed lines represent the naive and random baseline scores, respectively

Table 8 Confusion matrices for singular and plural nouns indicating how many of the correctly retrieved images contained only one or multiple visual referents to the target word. Here we show the counts at the penultimate phone and (parenthesized) the increase or decrease after having processed the final phone

Model	#refs in image	Noun morphology	
LSTM		singular	plural
	one	2470 (578)	3339 (−399)
	multiple	1851 (430)	2694 (187)
LSTM-VQ		singular	plural
	one	2374 (483)	3171 (−540)
	multiple	1704 (574)	2565 (189)

word is plural (although subtle prosodic cues might also be at play [64]). Figure 3 shows the P@10 scores from the gating experiment as a function of the gate number (number of phones processed so far), averaged over words of the same length. Unsurprisingly, recognition scores tend to increase as more phones are processed. Interestingly, for the plural nouns, recognition scores tend to drop at the last phone which, except for 'men' and 'women', is the plural suffix /z/ or /s/. The average P@10 value for plural target words drops from .517 to .479 between the penultimate and final gate for the LSTM model and from .513 to .449 for the LSTM-VQ model. It seems both VGS models have difficulty processing this suffix, the LSTM-VQ model even more so than the LSTM model.

A possible explanation for the P@10 drop is that, although the plural suffix causes the model to retrieve fewer images with single visual referents and more images with multiple referents (see Table 7), the decrease in single-referent images is greater than the increase in multiple-referent images. Table 8 shows the same confusion matrices as Table 7 but for the phone sequence up to the penultimate gate instead of the full word. The numbers between brackets indicate how the number of retrieved images changes upon processing the final phone. In case of plural nouns, the plural suffix is missing at the penultimate gate, so the model retrieves more images with a single referent, and fewer with a plural referent, than after also presenting the final phone. As can be seen in Table 8, and as hypothesised above, processing the plural suffix causes a drop in retrieval of single-referent images (−399) that is greater than the simultaneous increase in multiple-referent images (187), resulting in a drop in P@10 in Fig. 3.

Discussion

In this study we investigated the recognition of isolated nouns and verbs in a Visually Grounded Speech model. We were interested in whether visual grounding allows the

model to learn to recognise words as coherent linguistic units, even though our model is trained on full sentences and at no point receives explicit information about word boundaries or even that words exist at all. [45] used synthetic speech to test word recognition in their VGS model; we used newly recorded real speech. We could have opted to extract the words from spoken captions in the test set but this has a few disadvantages. Firstly, words in a sentence context are often significantly reduced and reduced word forms are hard to recognise in isolation even though they are perfectly recognisable in their original sentence context [65]. Secondly, due to co-articulation, we would not really be testing for single-word recognition unless the affected phones are removed, further reducing the word.

Word Recognition

Our first goal was to investigate whether the VGS model can recognise words in isolation after being trained on full utterances only. Our word recognition results show that our VGS model is able to recognise isolated target nouns. We have even shown that the LSTM model recognises both plural and singular nouns equally well even though plurals occur less often in the training data than singulars. While our scores are lower than those reported in [45], some difference was to be expected when working on real as opposed to synthetic speech. The average P@10 scores indicate that more than half of the top 10 retrieved images contain the visual referent and the models score well above the baselines. In fact, only four words (two in the LSTM model and two in the LSTM-VQ model) are not recognised at all, namely 'river' (in both models), 'ball' (LSTM) and 'waves' (LSTM-VQ). We saw that 'river' does return pictures of bodies of water (e.g. lakes or the ocean), and indeed it can be hard to discern the difference between a lake and a river from a picture. The fact that 'ball' is not recognised is a little baffling considering that 'basketball' has a P@10 score of .8 and 'football' a score of .4 (and pictures of either are also annotated as just 'ball').

We also tested whether models are able to recognise verbs in root, third person and participle form, the latter being the most common in the image descriptions. But even when we look only at the scores on the participle form, recognition scores for verbs are much lower than for nouns. In fact, most verbs are not recognised at all, and only 11 (LSTM) or 12 (LSTM-VQ) verbs have P@10 scores over .5. Looking at these words we see that many of them consistently occur together with an object (e.g. 'surfing', 'playing', 'skiing', 'holding' and 'racing') so the models might simply recognise the objects they co-occur with. This could be explained by our use of image features from ResNet-152, a network trained to recognise objects, not actions or body postures. However, it also recognises 'running', 'walking', 'jumping' and 'smiling', so the image features do seem to contain more information than simply the presence of a human in the image. Verb recognition in our model was far from good and this presents an interesting avenue for further research. We think it is possible for the VGS model to also learn to recognise actions, perhaps by fine-tuning parts of ResNet with the VGS model or training the visual side of the model from scratch like in [31].

Word Competition

In our gating experiment, we investigated whether the model's word recognition is affected by word competition, as is the case in humans. The results show clear evidence of word competition effects in our model. There is a strong effect of word-initial cohort size where recognition scores are lower when more words are possible given the current input sequence. We also find a positive effect of neighbourhood density that is modulated by a negative interaction with word count. This means that the effect of neighbourhood density is higher for lower-frequency words. This is in line with findings that, for humans, recognition of low-frequency words is facilitated by dense neighbourhoods whereas recognition of high-frequency words is facilitated by sparse neighbourhoods [50, 51].

We find a positive effect of neighbourhood density, contrary to what we may expect if we assume more word competition (i.e. a denser neighbourhood) makes word recognition harder. Furthermore, given the strength of the interaction with word count, the neighbourhood density effect is only negative for highly frequent words. [50] gives a possible explanation for the interaction between word count and neighbourhood density: during word learning, dense neighbourhoods have a positive effect on word recognition because hearing similar-sounding words facilitates learning. During word recognition, dense neighbourhoods have a negative effect because similar-sounding words compete for recognition. For infrequent words, the learning effect outweighs the competition effect, and vice versa. Our model may simply have been trained on too few of the most frequent words for the competition effect to outweigh the learning effect, explaining the overall positive effect of neighbourhood density. Together with the strong effect of initial cohort size, we argue that we do indeed see word competition effects in our VGS model.

Plurality

We also investigated whether our VGS model learns the difference between singular and plural nouns. Our results show that not only is the model able to recognise

target nouns in both forms but, to a limited extent, it also learns to differentiate between the two forms: when prompted with plural target nouns, the model retrieves more images with multiple referents and fewer with single referents than when prompted with single nouns (see Table 7). Thus, the model learns a meaningful difference between singular and plural nouns in terms of their visual representations.

P@10 scores from our gating experiment showed that words are recognised better when more of the word is processed. Yet, we also see that recognition scores are well above the baselines before word offset, which means that the model is able to recognise words from partial input. We take this to mean that the model not only recognises words, but is also able to encode useful sub-lexical information. However, at first glance, both models seemed to have trouble with the plural suffix. As shown by the results of the gating experiment, before the plural suffix recognition of plural target words is often more accurate than recognition of singulars. However, at the final phone, recognition scores of plural nouns drop and become equal or lower to that of singular nouns. While this seems to be evidence against the encoding of useful sub-lexical information, our results also show that presenting the model with plural nouns causes both models to retrieve *more* images with multiple visual referents and *fewer* images with a single referent. This indicates that the model encodes the plural suffix in a way that correctly affects recognition.

Using the recognition results from the gating experiment, we found that it is indeed only after the plural suffix that the distribution over single and multiple referents in the retrieved images shifts. At the gate just before the plural suffix (where the word is technically still singular), the model retrieves more single-referent images and fewer multiple-referent images than after the plural suffix. As previously said this is in contrast to human listeners, who are able to use subtle prosodic cues to recognise plural nouns [64]. It is not surprising that our current model, which is far from human performance in terms of word learning and recognition, is not able to exploit such cues, but this is an interesting avenue for further research.

Further analysis showed that after processing the plural suffix, the drop in single-referent images is larger than the increase in multiple-referent images. This may simply be caused by an imbalance in the test data; there are more annotations of single visual referents (3864) than multiple visual referents (2203). Further testing with a more balanced set of test images could show whether the performance drop seen in our gating experiment is indeed due to *correct* recognition of the plural suffix, as we would then expect the increase in retrieved multiple-referent images to outweigh the decrease in retrieved single-referent images.

Vector Quantisation

Our final research goal was to establish whether the addition of VQ layers to the VGS model aids in the discovery and recognition of words. Previous research had shown that VQ layers inserted into a VGS model learned a hierarchy of linguistic units; a phoneme-like inventory in the first layer, and a word-like inventory in the second layer [13]. VQ layers discretise otherwise continuous hidden representations by mapping neighbouring speech frames to the same embedding in the codebook. We expected that this aids in the discovery of words and perhaps even allows the LSTM-VQ model to recognise words earlier in the gating experiment, as the model is forced to output discrete units from its word-like VQ layer at every time step. Moreover, the codebook size (2048) is smaller than the total number of unique words in Flickr8k so, if anything, one would expect the model to prioritise highly frequent words, of which we took the top 50 as our targets.

In all of the experiments, however, we found no evidence of the VQ layers aiding in the recognition of words: we showed that the LSTM-VQ model slightly outperforms the LSTM model on the training task (image-caption retrieval) so it cannot be the case that it is simply not a good VGS model. With regard to word recognition performance, the LSTM-VQ model recognises singular nouns better than the LSTM model, but it performs much worse at recognising plural nouns. Also noticeable is a gap between singular versus plural noun recognition that is not present in the LSTM model (when looking at the subset of words that have both a plural and singular form).

Furthermore, both GLMMs showed no main effect of the presence of VQ layers on recognition scores. We did find a negative interaction between VQ and gate number, indicating that the effect of gate is smaller for the LSTM-VQ model than for the LSTM model. Considering that final recognition performance is similar between the two models, the smaller effect of gate means the LSTM-VQ model performs better at early gates. That is, it recognises words *earlier* than the LSTM model. Together, these results indicate that the addition of VQ layers is neither beneficial nor detrimental to word recognition performance, although the LSTM-VQ model requires less of the input sequence for correct recognition. An interesting question for future research is which model performs more 'human-like', that is, which model recognises words closest to the point where humans do.

Finally, we did a post hoc test for the interaction between VQ and morphology that shows the LSTM-VQ model has an advantage on the most frequent noun and verb forms, but performs worse on the less frequent forms. Perhaps this is due to the limited codebook size forcing the model to dedicate codes to the most frequent words in the training data.

Limitations

In this study, we trained and tested a model on real speech, as opposed to synthetic speech. As expected, overall recognition scores were lower than reported on synthetic speech, as natural speech is known to be more challenging for current models of speech recognition. However, the speech used in this study is read aloud speech, which is itself cleaner than spontaneous speech. In the interest of learning from data that is as natural as possible, spontaneous speech is preferred as this is the type of speech humans are most exposed to.

Furthermore, while we have shown that our model is capable of recognising words in isolation while only having seen those words in utterances, we selected only a small number of words. The small number mainly results from selecting only words with enough occurrences in the training data to reasonably expect the model to be able to learn to recognise the word, and enough occurrences of their visual referents in the test images in order to evaluate the recognition performance. On the other hand, given that the model was able to learn to recognise the words in this study after relatively little exposure, it is not unreasonable to expect the model to be able to learn more words if exposed to them.

Finally, our model depends on correlations between the speech signal and the images in order to learn to recognise meaningful constituents in utterances. Furthermore, our concept of 'recognition' of a word is defined as the retrieval of images containing its visual referent, limiting the model to 'visible' things, such as object nouns and action verbs (and not even all of those). As our results showed, the model especially struggles with verbs, even though we selected verbs with a visual referent (the actions referred to were definitely 'visible' as we were able to annotate their presence). As mentioned before, this may partly be due to the fact that we use a pre-trained *object* recognition network. However, it should be mentioned that the inter-annotator agreement for verbs was lower than for nouns, so even for the annotators, it was harder to determine the presence of actions than objects. We have argued here that visual information is an important learning signal in learning language; however, still images are but a single possible source of visual information. Actions can be partly defined by the movements involved, and as such, video might be a more appropriate learning signal.

Conclusion

We investigated whether VGS models learn to discover and recognise words from natural speech. Our results show that our models learn to recognise nouns. To a lesser extent, they are capable of recognising verbs but future research should look into the image recognition side of the model to further improve this. Our models even learned to encode meaningful sub-lexical information, enabling it to interpret the visual difference signalled by the plural morphology. Contrary to what we expected based on previous research, our results show no evidence that vector quantisation aids in the discovery and recognition of words in speech. Importantly, we investigated the cognitive plausibility of the model by testing whether word competition influences our models' word recognition performance, as we know happens in humans. We have shown that two well-known measures of word competition predict word recognition in our models and found evidence in favour of a disputed interaction between word count and neighbourhood density found in human word recognition.

Taking inspiration from human learning processes, our research has shown that using multiple streams of sensory information allows our model to discover and recognise words without any prior linguistic information from a relatively small dataset of scenes and spoken descriptions. Using realistic and naturally occurring input is important for creating speech recognition models that are more cognitively plausible, and visual grounding is an important step in that direction.

Funding The research presented here was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

Data Availability The datasets generated during and/or analysed during the current study are available through the following repositories: Flickr8k: https://forms.illinois.edu/sec/1713398. New data: https://doi.org/10.17026/dans-22n-xh47.

Code Availability All our code (model training, analysis) can be found on https://github.com/DannyMerkx/speech2image/tree/CogComp2022

Declarations

Ethics Approval All procedures performed in this study involving human participants were in accordance with the ethical standards of the Ethics Assessment Committee Humanities of the Radboud University Nijmegen, the Declaration of Helsinki and the ethics code of the American Psychological Association.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Conflict of Interest The authors declare no competing interests.

References

1. Benedict H. Early lexical development: Comprehension and production. J Child Lang. 1979;6(2).
2. Snyder LS, Bates E, Bretherton I. Content and context in early lexical development. J Child Lang. 1981;8(3).
3. Eisner F, McQueen JM. Speech perception. In: Stevens' handbook of experimental psychology, fourth edition. vol. 3 Language & thought. 4th ed. New Jersey: John Wiley; 2018. p. 1-47.
4. Weber A, Scharenborg O. Models of processing: lexicon. WIREs Cognit Sci. 2012:387-401.
5. Elman JL, McClelland JL. Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. J Mem Lang. 1988;27(2):143–65.
6. Marslen-Wilson WD. Functional parallelism in spoken word-recognition. Cognition. 1987;25(1):71-102. Special Issue Spoken Word Recognition.
7. Norris D. Shortlist: a connectionist model of continuous speech recognition. Cognition. 1994;52(3):189–234.
8. Norris D, McQueen J. Shortlist B: A bayesian model of continuous speech recognition. Psychol Rev. 2008;115:357–95.
9. Scharenborg O. Modeling the use of durational information in human spoken-word recognition. J Acoust Soc Am. 2010;127(6):3758–70.
10. ten Bosch L, Boves L, Ernestus M. DIANA, a process-oriented model of human auditory word recognition. Brain Sci. 2022;12(5).
11. Räsänen O, Rasilo H. A joint model of word segmentation and meaning acquisition through cross-situational learning. Psychol Rev. 2015;122(4):792.
12. De Deyne S, Navarro DJ, Collell G, Perfors A. Visual and affective multimodal models of word meaning in language and mind. Cogn Sci. 2021;45(1): e12922.
13. Harwath D, Hsu WN, Glass J. Learning hierarchical discrete linguistic units from visually-grounded speech. In: ICLR 2020 The Ninth International Conference on Learning Representations; 2020. p. 1-22.
14. Kamper H, Shakhnarovich G, Livescu K. Semantic speech retrieval with a visually grounded model of untranscribed speech. IEEE/ACM Transactions on Audio, Speech and Language Processing. 2019;27(1):89–98.
15. Roy D, Pentland A. Learning words from natural audio-visual input. In: 5th International Conference on Spoken Language Processing; 1998. p. 1279-82.
16. Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. J Artif Intell Res. 2013;47(1):853–99.
17. Chen X, Fang H, Lin TY, Vedantam R, Gupta S, Dollar P, et al. Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv: 1504.00325. 2015.
18. Merkx D, Frank SL. Learning semantic sentence representations from visually grounded language without lexical knowledge. Nat Lang Eng. 2019;25(4):451–66.
19. Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. p. 3128-37.
20. Klein B, Lev G, Sadeh G, Wolf L. Associating neural word embeddings with deep image representations using Fisher Vectors. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2015. p. 4437-46.
21. Ma L, Lu Z, Shang L, Li H. Multimodal convolutional neural networks for matching image and sentence. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE; 2015. p. 2623-31.
22. Vendrov I, Kiros R, Fidler S, Urtasun R. Order-embeddings of images and language. In: International Conference on Learning Representations (ICLR 2016); 2016. p. 1-12.
23. Wehrmann J, Mattjie A, Barros RC. Order embeddings and character-level convolutions for multimodal alignment. Pattern Recogn Lett. 2018;102:15–22.
24. Dong J, Li X, Snoek CGM. Predicting visual features from text for image and video caption retrieval. IEEE Trans Multimedia. 2018;20.
25. Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning. vol. 37; 2015. p. 169-76.
26. Harwath D, Glass J, Deep multimodal semantic embeddings for speech and images. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE. 2015;2015:237–44.
27. Harwath D, Torralba A, Glass J. Unsupervised learning of spoken language with visual context. In: Advances in Neural Information Processing Systems 29; 2016. p. 1858-66.
28. Chrupała G, Gelderloos L, Alishahi A. Representations of language in a model of visually grounded speech signal. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2017. p. 613-22.
29. Merkx D, Frank S, Ernestus M. Language learning using Speech to Image retrieval. In: INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association; 2019. p. 1841-5.
30. Havard W, Besacier L, Chevrot JP. Catplayinginthesnow: Impact of prior segmentation on a model of visually grounded speech. In: Proceedings of the 24th Conference on Computational Natural Language Learning. Association for Computational Linguistics; 2020. p. 291-301.
31. Harwath D, Recasens A, Surís D, Chuang G, Torralba A, Glass J. Jointly discovering visual objects and spoken words from raw sensory input. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 649-65.
32. Scharenborg O, Besacier L, Black A, Hasegawa-Johnson M, Metze F, Neubig G, et al. Speech Technology for Unwritten Languages. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020;28:964–75.
33. Kamper H, Roth M. Visually grounded cross-lingual keyword spotting in speech. The 6th Intl Workshop on Spoken Language Technologies for Under-Resourced Languages. 2018.
34. Kamper H, Shakhnarovich G, Livescu K. Semantic keyword spotting by learning from images and speech. arXiv preprint arXiv: 1710.01949. 2017.
35. Kamper H, Settle S, Shakhnarovich G, Livescu K. Visually grounded learning of keyword prediction from untranscribed speech. INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association. 2017:3677-81.
36. Wang X, Tian T, Zhu J, Scharenborg O. Learning fine-grained semantics in spoken language using visual grounding. In: Proceedings of the IEEE International Conference on Circuits and Systems; 2021. p. 1-5.
37. Srinivasan T, Sanabria R, Metze F, Elliott D. Fine-grained grounding for multimodal speech recognition. In: Findings of EMNLP 2020; 2020. p. 2667-77.
38. Palaskar S, Sanabria R, Metze F. End-to-end multimodal speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2018. p. 5774-8.
39. Chrupała G, Gelderloos L, Kádár Á, Alishahi A. On the difficulty of a distributional semantics of spoken language. In: Proceedings of the Society for Computation in Linguistics. vol. 2; 2018. p. 167-73.

40. Hsu WN, Harwath D, Glass J. Transfer learning from audio-visual grounding to speech recognition. In: INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association; 2019. p. 3242-6.
41. Chrupała G, Higy B, Alishahi A. Analyzing analytical methods: The case of phonology in neural models of spoken language. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. p. 4146-56.
42. Merkx D, Frank SL, Ernestus M. Semantic Sentence Similarity: Size does not Always Matter. In: INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association; 2021. p. 4393-7.
43. Räsänen O, Khorrami K. A computational model of early language acquisition from audiovisual experiences of young infants. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association. 2019:3594-8.
44. van den Oord A, Vinyals O, kavukcuoglu k. Neural discrete representation learning. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. p. 6306-15.
45. Havard WN, Chevrot JP, Besacier L. Word recognition, competition, and activation in a model of visually grounded speech. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). Association for Computational Linguistics; 2019. p. 339-48.
46. Scholten S, Merkx D, Scharenborg O. Learning to recognise words using visually grounded speech. In: Proceedings of the IEEE International Conference on Circuits and Systems. IEEE; 2021. p. 1-5.
47. Koch X, Janse E. Speech rate effects on the processing of conversational speech across the adult life span. J Acoust Soc Am. 2016;139(4).
48. Norris D, McQueen JM, Cutler A. Competition and segmentation in spoken-word recognition. J Exp Psychol Learn Mem Cogn. 1995;21(5):1209.
49. Luce PA, B PD. Recognizing spoken words: the neighborhood activation model. Ear and Hearing. 1998;19:1-36.
50. Metsala JL. An examination of word frequency and neighborhood density in the development of spoken-word recognition. Memory & Cognition. 1997;25(1):47–56.
51. Goh WD, Suáres L, Yap MJ, Tan SH. Distributional analyses in auditory lexical decision: Neighborhood density and word-frequency effects. Psychonomic Bulletin & Review. 2009;16(5):882–7.
52. Rispens J, Baker A, Duinmeijer I. Word recognition and nonword repetition in children with language disorders: The effects of neighborhood density, lexical frequency, and phonotactic probability. J Speech Lang Hear Res. 2015;58(1):78–92.
53. Garlock VM, Walley AC, Metsala JL. Age-of-Acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. J Mem Lang. 2001;45(3):468–92.
54. Cotton S, Grosjean F. The gating paradigm: A comparison of successive and individual presentation formats. Perception & Psychophysics. 1984;35(1):41–8.
55. Smith LN. Cyclical learning rates for training neural networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV); 2017. p. 464-72.
56. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770-8.
57. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009. p. 248-55.
58. Bengio Y, Léonard N, Courville CA. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. arXiv preprint arXiv: 1308.3432. 2013.
59. Kingma DP, Ba J. Adam: A Method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR); 2015. p. 1-15.
60. van Niekerk B, Nortje L, Kamper H. Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge. In: INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association; 2020. p. 4836-40.
61. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal. 2017;9(2):378–400.
62. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Toolkit The Kaldi Speech Recognition, In: IEEE, et al. Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society. 2011;2011:1–4.
63. Vitevitch MS, Luce PA. Phonological neighborhood effects in spoken word perception and production. Annual Review of Linguistics. 2016;2:75–94.
64. Kemps RJJK, Ernestus M, Schreuder R, Baayen RH. Prosodic cues for morphological complexity: The case of Dutch plural nouns. Mem Cogn. 2005;33:430–46.
65. Ernestus M, Baayen H, Schreuder R. The recognition of reduced word forms. Brain Lang. 2002;81:162–73.