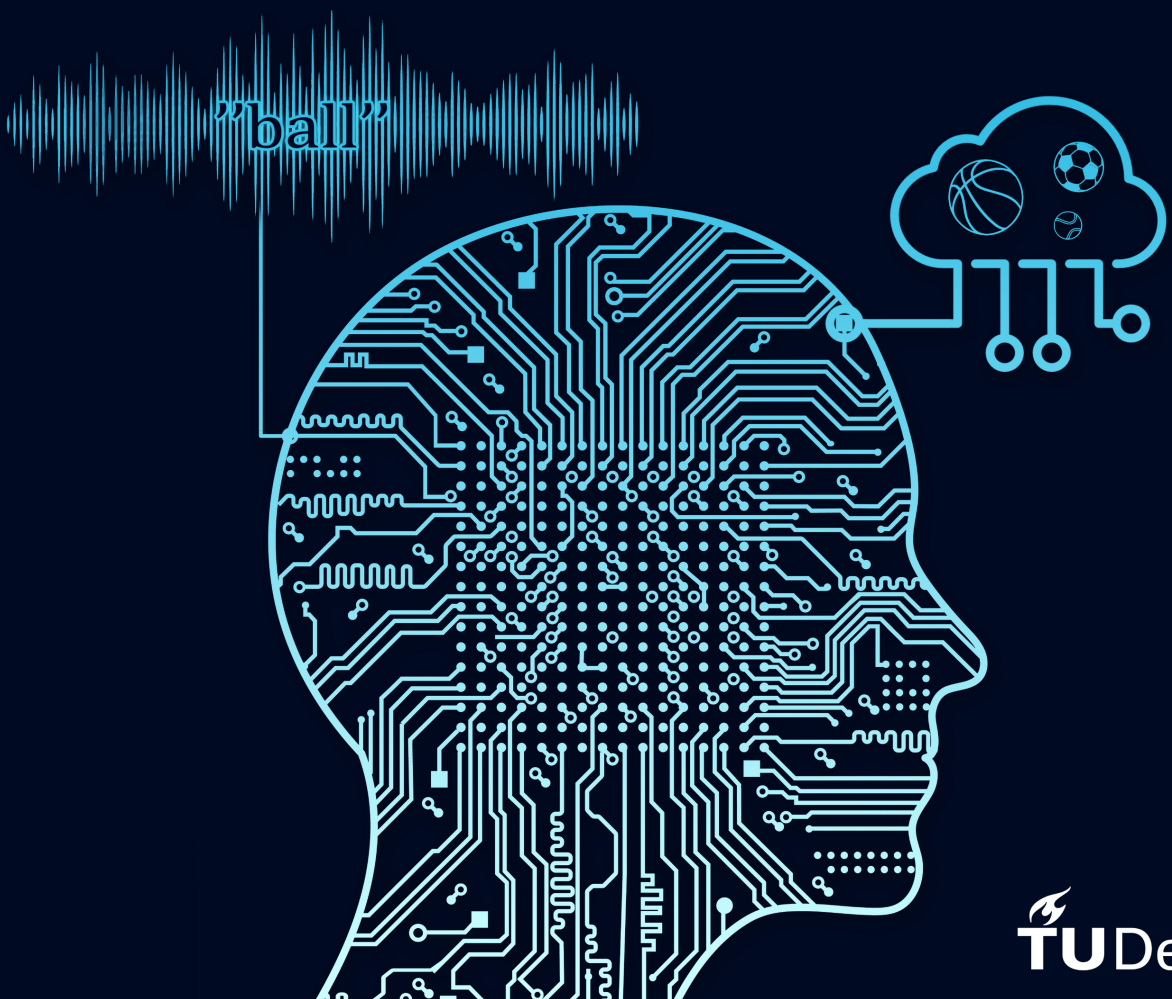
WORD RECOGNITION IN A MODEL OF VISUALLY GROUNDED SPEECH

by Sebastiaan Scholten



TU Delft

WORD RECOGNITION IN A MODEL OF VISUALLY GROUNDED SPEECH

AN ANALYSIS USING TECHNIQUES INSPIRED BY HUMAN SPEECH PROCESSING RESEARCH

Master Thesis

to obtain the degree of Master of Science
at Delft University of Technology,
to be defended publicly on Friday July 24th, 2020 at 13:00PM

by

SEBASTIAAN SCHOLTEN

Course:	Master Computer Science	
Student number:	4690591	
Thesis committee:	Dr. Odette Scharenborg,	TU Delft, supervisor
	Dr. Nava Tintarev,	TU Delft
	Dr. Catharine Oertel,	TU Delft
	Danny Merkx MSc,	Radboud University, supervisor

TU Delft

ABSTRACT

A Visually Grounded Speech model is a neural model which is trained to embed image-caption pairs closely together in a common embedding space. As a result, such a model can retrieve semantically related images given a speech caption and vice versa. The purpose of this research is to investigate whether and how a Visually Grounded Speech model can recognise individual words. Literature on Word Recognition in humans, Automatic Speech Recognition and Visually Grounded Speech models was evaluated. Techniques used to analyse human speech processing, such as gating and priming, were taken as inspiration for the design of the experiments used in this thesis.

Multiple aspects of words recognition were investigated through three experiments. Firstly, it was investigated whether the model can recognise individual words. Secondly, it was investigated whether the model can recognise words from a partial sequence of its phonemes. Thirdly, it was investigated how word recognition is affected by contextual information. The experiments show that the model is able to recognise words while not being supervised for that task, and that factors such as word frequency, the length of a word and the speaking rate affect word recognition. Furthermore, the experiments reveal that words can be recognised from a partial input of a word's phoneme sequence as well, and that recognition is negatively influenced by word competition from the word initial cohort. Furthermore, the word recognition in context experiment reveals that contextual information can enhance the recognition of words which are recognised less well.

PREFACE

AFTER months of writing and thousands of lines of code, I have been eagerly looking forward to handing in my thesis and obtaining the coveted Master of Computer Science degree. However, it also fills me with regret to have to leave behind this wonderful place of knowledge and challenge at TU Delft.

This thesis trajectory has confirmed once more for me what I love the very best about computer science: the limitless possibilities, the delving deeper, experimenting on uncharted territory, the concrete results and the potential impact on science, business and society through the development of technologies which have the power to bring about change.

Although people outside of the CS field may not understand every detail of this research, it is about something everybody is familiar with: the process of learning a language. It has been an exciting challenge to try and discover whether computers can be made to learn languages in a way similar to the way humans do: i.e. through learning words by being exposed to visual stimuli and speech signals.

This research would not have been possible without the state-of-the-art research by other members within the global CS community to build upon and the invaluable support of some extremely knowledgeable people within the computer science departments of TU Delft and other universities in the Netherlands. I particularly would like to thank Odette Scharenborg and Danny Merkx, who supervised me through the process of writing my thesis. Without the video calls and countless emails sent back-and-forth, it would have been impossible to bring this thesis to where it is at now. Their knowledge and encouragement have opened up this exciting path for me and firmly planted the seed for my future ambitions in this field. Also, with their help, I have submitted a paper containing preliminary results to Interspeech 2020 (see Appendix-A). I would also like to express my gratitude to the other members of this thesis committee, Nava Tintarev and Catharine Oertel, who agreed without reserve to dedicate hours of their time to the evaluation and defense of this thesis. Also, I would like to thank the participants of our weekly SALT (Speech And Language Technology) group meetings for their valuable feedback and advice throughout the past months.

Sebastiaan Scholten
Delft, July 2020

CONTENTS

1

INTRODUCTION

1

THE development and improvement of Automatic Speech Recognition (ASR) has been one of the greatest successes in the field of speech technology over the past several years. Speech technology has found its practical use in a range of different settings, such as virtual assistants, messaging apps, search engines, and home automation. However, for an ASR system to recognise words from spoken speech, a large vocabulary needs to be learned in a supervised setting, which requires a large amount of transcribed speech data. For many low-resource or unwritten languages, such transcribed training data is not readily available, which makes this technology inaccessible for those languages.

Humans, on the other hand, are able to learn vocabularies from raw sensory input in notably more difficult scenarios. It is theorised that the fact that babies repeatedly hear certain words while they observe certain objects around them, enables them to learn a mapping between speech and objects [1]. Repetitive hearing of these utterances in the context of some functional consistency, such as picking up an object, displays the meaning of a smaller constituent of such an utterance, e.g., a word, and potentially the class of objects it belongs to [2].

Some core principles of Visually Grounded Speech (VGS) models are inspired by these natural learning processes. While ASR models use speech signals only, Visually Grounded Speech models include visual information instead of textual transcriptions to guide the training of models [3]. This multi-modal approach is more closely inspired by human language learning, and provides the advantage of removing the need for transcribed training data. Recently, there has been an increasing interest in learning languages through such multi-modal approaches, which directly pair speech signals with visual input [3–10].

VGS models can have a big influence on speech technologies in the short and long-term. In the short-term, speech technologies such as retrieving images given a speech signal could be developed for low-resource or unwritten languages. In the long-term, VGS models could potentially be used to develop speech recognition models, or be used in robotics, where such models could learn languages based on co-occurring audio and video signals [11].

Currently, we are still far away from developing VGS models as speech recognisers which can directly map speech to textual transcriptions. Traditional ASR systems are trained on transcribed segmented words specifically for the speech recognition task, as opposed to VGS models which are trained on images and speech for retrieving images given speech. In order to determine whether and how VGS models can be used in speech technologies, more needs to be learned about whether a VGS model actually learns to *recognise words* from the images and speech it is trained on. Furthermore, little is known about the inner workings of a VGS neural model, in particular about how individual words are activated and stored in the neural model.

In order to explore and add to the potential development of such future speech technologies, the purpose of this research is to investigate whether and how a VGS model performs word recognition. In the following section, the problem statement, and the research questions that follow from it, will be discussed.

1.1. PROBLEM STATEMENT AND RESEARCH QUESTIONS

VGS models are not explicitly trained to work as speech recognisers, but rather to create visual-semantic alignments between images and speech signals. VGS models are trained on images with relevant speech captions, in order to embed them in a common embedding space. As a result these models can retrieve an image embedding when given a speech embedding, and vice versa. Since the VGS model does not return a textual transcription when given a word, *word recognition* can be evaluated by first embedding images and words, and then retrieving image embeddings with a high cosine similarity to the word embeddings. The textual captions of the images can then serve as ground-truth labels for evaluating whether the visual referent of a word was present in the image, and therefore whether the word was 'recognised'.

Although VGS models are not explicitly trained as speech recognisers, recent research indicates that VGS models learn to recognise meaningful sentence constituents such as phonemes (units of sound in linguistics) and words from the speech captions it is trained on [5, 6, 12–14]. It appears that a VGS model does not just encode these constituents into the speech embeddings, but that the model actually 'recognises' individual words and learns to map them onto their correct visual referents. This means that the VGS model implicitly learns phonemes and words from the speech captions it is trained on, as opposed to learning a vague acoustic representation of a speech signal.

Moreover, research by Havard and colleagues showed that their VGS model was able to reliably map individual words to their visual referents, which indicates that a VGS model is able to perform word recognition [12]. Havard and colleagues used synthetically generated speech captions, while in this research word recognition will be investigated with the Flickr8k dataset, which contains naturally spoken speech captions. Word recognition is expected to be more challenging with naturally spoken speech as opposed to synthetic speech, due to naturally spoken speech having more variation in quality, noise and speaking rate than synthetic speech.

Therefore, the first aim of this research is to investigate whether a VGS model can recognise words based on naturally spoken speech signals. Consequently, the first research question is formulated as follows:

- Research Question 1: *Can a Visually Grounded Speech model perform word recognition with naturally spoken speech?*

So far, little is known about the inner workings of a VGS neural model, in particular about *how* a VGS model recognises words based on naturally spoken speech. Therefore, the second research question is defined as follows:

- Research Question 2: *How are naturally spoken words recognised by a Visually Grounded Speech model?*

Taking inspiration from human speech recognition (HSR) research, this question will be approached from several different perspectives, which will be addressed through five sub research questions.

Firstly, word recognition can be investigated from the perspective of word activation. Word activation in humans refers to the 'goodness' of fit between a sensory input

and the mental representation of a word [15]. As a result, there can be different levels of activation of the word in the mental representation of a human, based on how much word information is heard. This research will *investigate word recognition in a VGS model through looking at word activation*. A gating experiment, inspired by HSR, will be performed where words, segmented at increasing length from word onset and offset, are presented to the model in order to determine the amount of word activation at different points in time. Looking at these word activations could give insight into the 1) time-course of word recognition, 2) the amount of information needed for word recognition, and 3) whether the neural model is able to encode units of sound, such as phonemes.

- Sub question 1: *What is the time-course of word recognition?*
- Sub question 2: *What is the amount of information needed for word recognition?*
- Sub question 3: *Is the model able to encode units of sound, such as words and phonemes?*

Secondly, it is well known that contextual information, such as for example a part of the sentence surrounding a word, can aid word recognition in humans if provided as sensory input [16]. In HSR this can be investigated through a priming experiment, where a human listener is provided with both primed and unprimed sensory input. In this research a priming experiment, inspired by HSR, will be performed, where the VGS model's word recognition ability is evaluated by comparing its performance in a 'primed' setting (with contextual information) with its performance in an 'unprimed' setting (without contextual information). This experiment can reveal if, like in humans, *contextual information can aid or hinder word recognition* in a VGS model.

- Sub question 4: *How does contextual information affect word recognition?*

Lastly, this research aims to uncover *acoustic and linguistic factors which inhibit or aid word recognition*. Psycholinguists and ASR researchers have been interested in the cross-fertilisation of HSR and ASR research for quite some time [17]. Following these efforts, this research aims to uncover whether factors affecting word recognition in humans also affect word recognition in VGS models.

- Sub question 5: *What linguistic and acoustic factors affect word recognition?*

The rest of this thesis is organised as follows. Firstly, relevant literature on Word Recognition in humans (Section 2.1), Automatic Speech Recognition (Section 2.2) and Visually Grounded Speech models (Section 2.3) will be discussed. Secondly, the methodology (Section 3) behind the experiments which were performed to answer the research questions will be explained. Thirdly, the results (Section 4) of these experiments will be presented, followed by a discussion of the results together with the limitations of this research (Section 5). Afterwards, several recommendations for future research will be outlined (Section 6). Lastly, the research questions will be answered in the conclusion, followed by a conclusive statement regarding this work's contributions (Section 7)

2

RELATED WORK

THE purpose of this research is to analyse whether and how a VGS model performs word recognition. In order to investigate this, this chapter presents a review on three distinct topics relevant to this research:

2

1. Word Recognition in humans
2. Automatic Speech Recognition
3. Visually Grounded Speech models

2.1. WORD RECOGNITION IN HUMANS

The process of recognising words is a seemingly effortless practice for humans, however the underlying process is very complex, and constitutes an active research topic for psycholinguists. Whenever humans engage in language use, they map incoming auditory information onto their mental lexicon, which contains the words they know. The sounds, which are called 'phones' in linguistics, that best resemble the incoming speech signal are 'activated'. These activated phone representations, activate every possible word in which they appear, irrespective of the position of the phone in the word. As more speech information becomes available, words that no longer match the input will drop out of the list of activated words. Words that are activated are called 'competitors' or 'competitor words'. The competitor word that best matches the speech input is recognised, in a process called word recognition [18].

In order to model this behaviour of accessing words in our mental lexicon, psycholinguists have developed models of spoken-word recognition. The purpose of these models of spoken-word recognition is to explain the different stages and mechanisms which are at play during word recognition in humans [18]. More recent models of spoken-word recognition can be implemented as computer programs, which allows testing of the 'goodness' of fit between a word recognition theory and collected behavioural data [19]. A number of these spoken-word recognition models will be discussed in the following section.

2.1.1. MODELS OF SPOKEN WORD RECOGNITION

The first of these spoken-word recognition models, which was introduced in the 1980s, is the COHORT model [20]. According to the COHORT model, a speech signal would activate all words with the same word onset (i.e. start of the word) in the listener's mental lexicon, and as more speech information becomes available to the human listener, words that no longer match the input are progressively ruled out. As a result, a single word in a listener's mental lexicon would be left which matched the speech signal, and the word would be recognised. This spoken-word recognition model had a number of shortcomings, namely that it did not incorporate that human listeners can recognise words that mismatch from the word onset [21], and that human listeners recognise words occurring more frequently in a language more easily [22]. However, this model sparked the development of new models of spoken-word recognition which did account for such intricacies. An example of one of these is the TRACE model of spoken-word recognition [23], which has been computationally implemented. The TRACE model accounted for some

shortcomings of the COHORT model, for example the consideration of word frequency, which was introduced by Dahan and colleagues [24], and that words can be activated for any part of the speech signal and not only with the word onset available to the human listener [18]. However, a significant drawback of this model remains, namely that TRACE can only use very small lexicons, so in order for it to realistically represent human word recognition, it would have to be able to use larger lexicons [18]. In order to overcome this problem, SHORTLIST was introduced [25], which is a two-staged model that allows a more realistically sized lexicon [18]. There are many more models of spoken-word recognition, which fall outside the scope of this research to cover. For a more detailed summary please refer to [18].

Psycholinguists have not reached a consensus as to which model best captures human word recognition, and it is still an active field of discussion and research. However, there are a number of mechanisms on which they have reached a general consensus. The first of those is that multiple word candidates are activated in the mental lexicon when a word or part of a word is being heard [26]. The second is that these word candidates are competing to be recognised [18].

2.1.2. FACTORS AFFECTING WORD RECOGNITION

There are a number of different linguistic and acoustic factors which are known to affect word recognition in humans. In the following paragraphs, a number of factors which are also considered in the experiments in this research will be discussed.

A factor known to affect word recognition in humans is word frequency. Word frequency refers to how frequently a word occurs in a language [22]. Multiple researchers have found word frequency to affect word recognition, where more frequent words are recognised faster, and with lower error rates than words which appear less frequently in a language [22, 27].

Furthermore, due to the large variability in the speech signal [18], the manner in which a speech signal is uttered can have an effect on how well it can be recognised. [18] For example, a factor known to affect word recognition is the speaking rate of the speaker. The speaking rate can be calculated by dividing the amount of phonemes present in a word by the length of a speech signal. For example, research by Sommers and colleagues found that human listeners could identify words in a speech signal less well when they were presented with multiple words at mixed speaking rates, as opposed to being presented with multiple words at a single speaking rate [28].

Furthermore, vowels and consonants in a speech signal appear to play a different role of importance in the process of word recognition. Research by Cole and colleagues investigated the relative importance of vowels and consonants by replacing vowels or consonants in words with noise, and asking human listeners to try and transcribe the words they heard [29]. They found that word recognition was considerably more reliant on vowels than on consonants, because twice as many words were recognised if vowels were retained in the speech signal than if consonants were retained [29].

Also, the number of competitor words plays a role in human speech processing: the more competitors there are, the longer it generally takes for a word to be recognised [30]. A way to represent the number of words competing for word recognition is by calculating the size of the word-initial cohort [31]. The word-initial cohort stems from the

original COHORT theory. When the first phonemes of a word are heard, a set of words is activated in the mental lexicon which share an initial sequence of phonemes, the word-initial cohort. As progressively more phonemes of a word are heard, the word-initial cohort diminishes in size as less competitor words match the initial word onset [20]. Gating experiments have shown that humans subconsciously produce such sets during word recognition [31, 32], and that the size of the word-initial cohort appears to affect word recognition.

Another way to quantify the amount of competitor words that are activated during word recognition is the phonological neighbourhood density. The phonological neighbourhood density refers to the number of words that differ a single phoneme from the target word [33]. The phonological neighbourhood density thus represents a set of words which are phonologically very similar to a word, and are likely to be activated as competitor words during word recognition.

2.1.3. GATING AND PRIMING

The process of word recognition can be researched through a method called 'gating'. Gating experiments are used to investigate the relation between the acoustic signal and access to the mental lexicon. In a gating experiment, words are repeatedly presented to a subject, with increasing length from word onset on each successive pass [34], allowing researchers to investigate the amount of input required for a subject to correctly recognise a word. Furthermore, gating allows the time-course of word recognition to be evaluated at each 'gate', which can reflect how strongly words are activated at each time step within a word [35]. Through a gating experiment, Grosjean and colleagues showed that not all words are recognised before their acoustic offset, showing that word recognition does not have to be strictly sequential [36]. This raised some problems for spoken-word recognition models such as the COHORT model, which works on the premise that words are recognised sequentially.

The process of storing and accessing words in the mental lexicon can be researched through priming experiments. Priming experiments aim to investigate whether exposure to a certain stimulus, a prime, can facilitate or inhibit the speed of recognition of a subsequent stimulus, i.e. the 'target word' [16]. It is well known that recognition of words by humans can be activated or suppressed by priming effects, thus hindering or aiding the speed of recognition [37]. For example, semantic priming experiments have shown that presenting a human listener with a semantically related prime, such as 'cat', can help with speed of recognition of the word 'dog' in comparison to first presenting the listener with an unrelated prime such as 'book' [38]. As a result, these gating and priming experiments can provide insight into how words are activated in a human listener's mental lexicon.

2.2. AUTOMATIC SPEECH RECOGNITION

Automatic Speech Recognition (ASR) technology has become ever so prevalent in our everyday lives. For example voice assistants, which have long been regarded as science fiction, have quickly become a familiar part of many services used on a daily basis. There are two main components in an ASR system, namely an acoustic model, which represents the relationship between an acoustic signal and some linguistic unit e.g. a phoneme, and a language model which models the statistical relationship between words in a language. Until about a decade ago, most ASR systems were based on Hidden Markov Models (HMMs) for language modelling together with Gaussian Mixture Models (GMMs) for building acoustic models [39]. In the past decade however, Deep Learning models have taken speech recognition by storm, and have become the go to approach for acoustic modelling [40, 41], helping to lower word error rates to as low as 4.9% on English continuous speech recognition tasks [42]. The next section will further elaborate on the topic of Deep Learning and how it is applied in speech recognition tasks. After that, a number of acoustic features which are used for ASR tasks will be discussed.

2.2.1. DEEP NEURAL NETWORKS

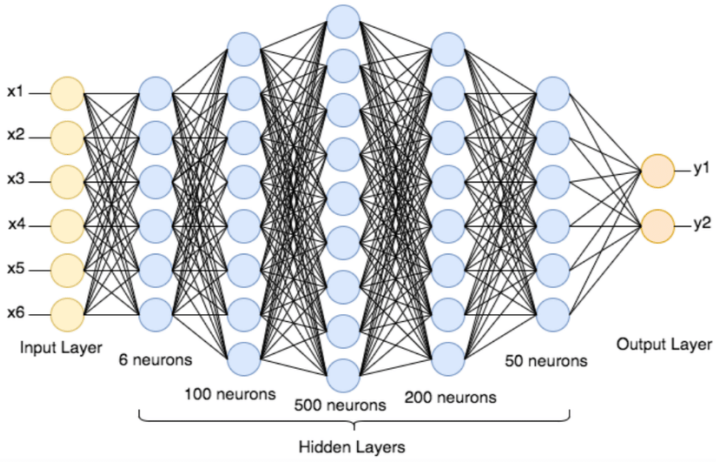


Figure 2.1: A Deep Artificial Neural Network containing multiple layers of neurons stacked on top of each other.

Deep Learning is a subfield of Machine Learning that aims to tackle machine learning tasks through the use of Artificial Neural Networks (ANNs). As displayed in Figure 2.1, an ANN is a mathematical model of an interconnected group of neurons that is loosely inspired by the neuron activations that happen in a biological brain. Neurons are connected through weights which contain activation functions to transform the input which passes through the connections, and the weights in those connections can be adjusted in order for the model to make predictions based on its input. As seen in Figure 2.1, an ANN contains multiple layers of connected neurons, called hidden layers, where parameters can be passed through.

In a process called supervised learning, an ANN can be given a set of desired outputs

for an input, and it adjusts the weights in the connections to transform its input to the desired output. This requires a loss function which can be used for calculating the distance between the *computed* output and the *desired* output. This loss can be propagated through the ANN weights through an algorithm called backpropagation, in order for the network to adjust its weights to minimize its loss, thus bringing the desired output closer to the computed output. As a result, an ANN can learn to transform a given input to a desired output for classification or regression problems.

ANNs started to perform really well on machine learning tasks after researchers started to implement deep architectures, containing many layers of neurons stacked on top of each other (as shown in Figure 2.1). These Deep Neural Networks (DNNs) boasted many more learnable parameters, allowing complex patterns to be learned from data. This allowed DNNs to be applied to large scale learning problems [43], such as speech recognition, where they have to be trained on large amounts of data in order to make accurate predictions. When DNNs started to be applied to speech recognition tasks, they quickly surpassed traditional GMMs in terms of word error rates [44]. In the next section, it will be discussed how DNNs are used in ASR, and which specific DNN architectures are often applied. Afterwards, the acoustic features used for training such DNNs will be discussed.

2.2.2. ASR WITH DEEP NEURAL NETWORKS

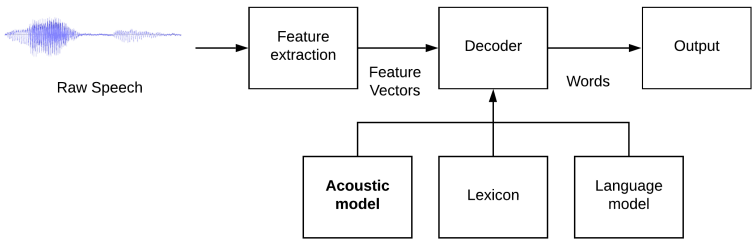


Figure 2.2: A visual representation of an ASR pipeline using an acoustic model, lexicon and language model in order to perform a word prediction based on acoustic features extracted from raw speech.

In Figure 2.2, a pipeline for an ASR system is shown. Acoustic features are computed on speech signals, in order to extract a vector of features to represent each frame within a speech signal. The decoding of word content within a speech signal is performed simultaneously by an acoustic model, a lexicon and a language model. An acoustic model is trained to recognise units of sound, e.g. phonemes, present within a speech signal. The lexicon is a corpus which stores how words are pronounced phonetically, and can be used to transform the phoneme output from the acoustic model into words. The language model is trained on large amounts of text and represents the probabilistic properties of a language, which allows it to make predictions as to which words will most likely follow from a sequence of words.

DNNs are often used for the task of acoustic modelling in combination with an HMM for language modelling. DNNs can also be used to perform ASR in an end-to-end configuration, where a single DNN is used to directly map acoustic features to words [45].

Two specific DNN architectures will be discussed, which - next to the traditional DNN architecture introduced in Section 2.2.1 - helped to achieve the low word error rates seen in speech recognition systems today. The two DNN architectures are Convolutional Neural Networks (CNNs) [46] and Recurrent Neural Networks (RNNs) [47]. These DNN architectures can be trained on transcribed speech data, while working as an acoustic model or as an end-to-end configuration, and as a result can find phonemes or words that are present in a speech signal. The general configuration of these DNNs, and their particular applicability to ASR will be discussed in the following paragraphs.

CONVOLUTIONAL NEURAL NETWORKS

Different from the traditional hidden layers in a DNN as shown in Figure 2.1, a CNN uses a specialized linear operator to transform the input passed through the neural network. As shown in Figure 2.3, a CNN uses convolution and pooling layers, where the convolution layer 'convolves' (see [48] for more information on the convolution operation) a matrix of weights (i.e. a kernel) over an input, such as an acoustic feature, and the pooling layer subsamples information from the input in order to retain the most important information [49]. As shown in Figure 2.3, these convolution and pooling layers can be stacked on top of each other, allowing the dimensionality of the input data to be reduced while extracting key describing features. Furthermore, the weights in the kernel, which are convolved over the input data, can be trained to discover patterns in the data, allowing CNNs to learn to transform the input to a desired output.

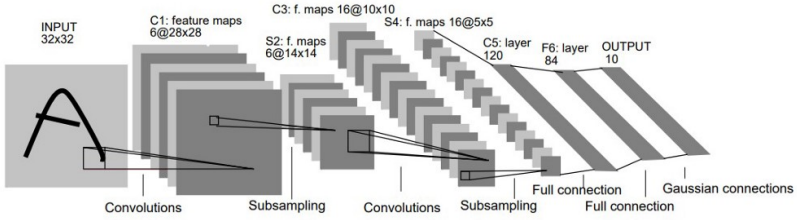


Figure 2.3: A Convolutional Neural Network architecture introduced by [46] showing how an input image is passed through multiple convolution and pooling layers in order to produce an output.

Particularly interesting about a CNN is that - other than a traditional DNN as shown in Figure 2.1 - it can process 2 or 3-dimensional input data while retaining the spatial structure of data. As a result, CNNs are often applied in image recognition tasks [46], but CNNs can also be used on 2 or 3-dimensionally organised acoustic features, and as a result be applied to speech recognition tasks [40]. Using CNNs for speech recognition is desirable due to three inherent properties: locality, parameter sharing, and pooling [50]. These properties allow a CNN to show a certain invariance to changes in acoustic features, i.e. changes created by small amounts noise and between speaker differences, which are beneficial to neglect when performing ASR [40].

2

RECURRENT NEURAL NETWORKS

RNNs use a specialized architecture that facilitates the modelling of sequential data [41]. Given that speech is also a temporal sequence of data, an RNN should inherently be able to model speech sequences well. An RNN differs from a traditional ANN as shown in Figure 2.1 by having sequential connections between neurons inside a single hidden layer, as shown in Figure 2.4. This mechanism facilitates the modelling of temporal sequences, because activations from earlier neurons in a sequence can be carried over to neurons further down the sequence within the same hidden layer. As shown in Figure 2.4, each neuron's activation is calculated as a function of the weighted sum of activations of the neurons located before it in the sequence.

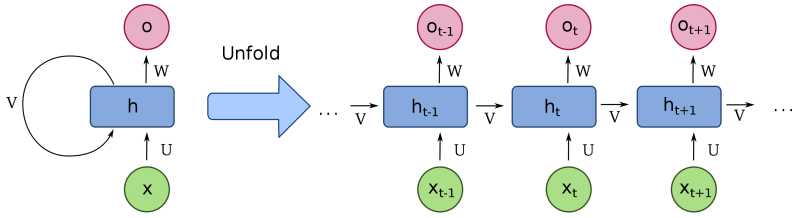


Figure 2.4: A generalized display of the recurrent connections between neurons in a hidden layer in a basic Recurrent Neural Network.[1]

Initially, RNNs were difficult to train due to the vanishing gradient problem, where gradients become vanishingly small during backpropagation, which made it difficult to train the weights in the recurrent connections between neurons [51]. This problem was overcome by the introduction of gating mechanisms such as Long-Short Term Memory (LSTM)[51] and Gated-Recurrent Units (GRU)[52]. These gating mechanisms replaced the recurrent connections between neurons in a hidden layer with gating units such as the GRU shown in Figure 2.5.

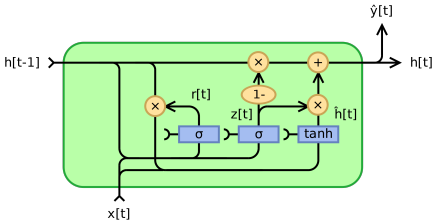


Figure 2.5: A GRU with reset gate r_t to forget certain information from its input, and update gate z_t to determine how much information from its input it wants to keep.[2]

[1] https://commons.wikimedia.org/wiki/File:Recurrent_neural_network_unfold.svg#/media/File:Recurrent_neural_network_unfold.svg

[2] https://upload.wikimedia.org/wikipedia/commons/3/37/Gated_Recurrent_Unit2C_base_type.svg

As shown in Figure 2.5, these GRUs contain reset and update gates r_t and z_t which allow the GRU to determine what part of the information, that passes through, it wants to keep, and what part it wants to forget. These reset and update gates help avoid gradients from becoming vanishingly small during backpropagation, as these gates essentially allow neurons to be 'skipped' when calculating the gradient. Stacking multiple layers of neurons with gating units, similar to stacking multiple layers as shown in Figure 2.1, allows the network to learn long-range time dependencies within its input data, which is beneficial when modelling temporal information such as speech [41].

In order for an RNN to make predictions from an acoustic feature to text (as is done in ASR), an encoder-decoder architecture is used to make a sequence-to-sequence prediction. A sequence-to-sequence model tries to map a fixed length input with a fixed length output where the length between an input and output may be different [53]. In an encoder-decoder architecture, the encoder creates an encoding vector to best represent the information in the input acoustic features. Consequently, the decoder receives this encoding vector as input and decodes the sequence of text from it. In order for an RNN to determine which part of the input sequence it should focus on, an attention-mechanism can be used [54]. An attention-mechanism ranks the relevancy of certain parts of the input received from the encoder, and can therefore decide which parts of the input sequence it should focus on in its decoding [55]. An attention-mechanism in an RNN allows the model to focus on certain frames of the input sequence when predicting certain frames of an output sequence, which enhances the quality of learning through the discarding of less important information in an acoustic feature [56].

2.2.3. ACOUSTIC FEATURES

To train neural models for speech recognition, speech signals are often transformed into an acoustic feature. The purpose of an acoustic feature is to represent acoustic properties of a speech signal in such a way that it can be given as input to an ANN [57]. Furthermore, an acoustic feature can remove information which is less relevant for speech recognition from a speech signal, such as power, pitch and vocal tract configuration [58].

Generally the process of developing acoustic features is divided into three stages. Firstly, the spectral envelope of the power spectrum is calculated, which is the curve outlining the distribution of power inside the frequency components of a speech signal at different time steps. Secondly, a number of features are calculated, differing per acoustic feature method. Lastly, these calculated features are compressed into a smaller set of features in order to arrive at the final representation of a speech signal at different time steps [59].

Mel-Frequency Cepstral Coefficients (MFCCs) are arguably the most well-known acoustic feature for training speech recognition models. MFCCs were introduced by [60], and have since been widely applied in ASR models with state-of-the-art performance. To create MFCCs, the power spectrum over frames of the signal is calculated using the fast-fourier transform. Then, logarithmic Mel-scale filterbanks are applied to all the transformed frames. Lastly, the discrete cosine transform is calculated over the logarithmic Mel spectrum on all frames, resulting in the final MFCC representation for a speech signal. Through applying the logarithmic Mel-scale filterbanks this technique better approximates characteristics of the human auditory system than linearly-space frequency

bands [61]. The coefficients are robust and have shown to be reliable when being used with different speakers and recording conditions [59], which makes this acoustic feature suitable for speech recognition tasks. Generally, 10-13 of the MFCC coefficients are used for speech recognition [62], however often researchers expand this by adding first-order (Δ) and second-order variations ($\Delta\Delta$) of the computed MFCCs [40], as this has often shown to improve speech recognition performance [63].

Linear predictive coefficients (LPC) are also applied as an acoustic feature for speech recognition. The premise of LPC is that a speech sample at a specific time can be approximated as a linear combination of past speech samples [57]. Then, by minimising the sum of squared differences between the actual sample and the predicted ones, a signal can be uniquely represented through a set of derived coefficients. The result is a signal, compressed to a unique set of coefficients, which is very memory efficient. In some research, LPC is criticized for losing vital information in a speech signal due to it being a linear approximation of a speech signal, while human speech is inherently non-linear [59]. However, it has achieved successes in some speech recognition tasks, such as in embedded systems in robotics, where memory-efficiency is a priority [64].

2.3. VISUALLY GROUNDED SPEECH MODELS

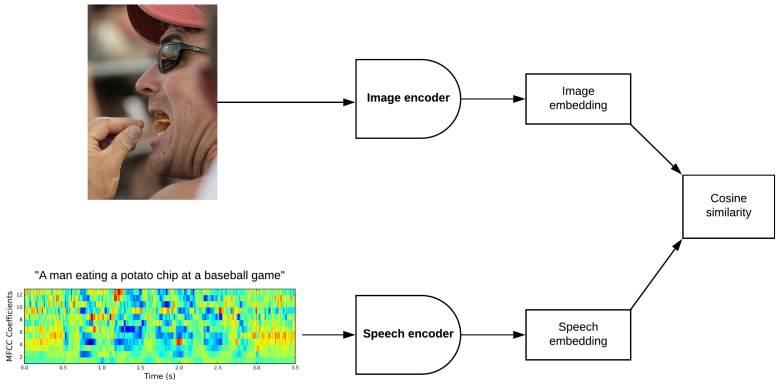


Figure 2.6: A simplified representation of the interplay between the speech and image encoders in the dual-encoder structure of a Visually Grounded Speech model.

The traditional ASR model architectures described in Section 2.2.1 often require substantial amounts of transcribed training data to accurately recognise words from speech signals. Recently, there has been an increasing interest in building speech technologies which directly align speech signals to images [3–9]. These Visually Grounded Speech (VGS) models use visual information rather than textual transcriptions to guide the training of the model. Following the approach of multimodal neural models which produce visual-semantic alignments for images and text [65], a VGS model employs two parallel Deep Neural Networks (DNNs) which can be trained to map a speech signal and a corresponding image into a common embedding space.

Figure 2.6 shows a simplified representation of a VGS model, where an image encoder and speech encoder can be used to create image and speech embeddings. If the model is provided with semantically related image-caption pairs, the model can be trained to minimise the cosine distance between these pairs. As a result, speech captions and images that are semantically similar are embedded closely together, i.e. with high cosine similarity, and semantically unrelated speech captions and images are not. A speech embedding can then be used to retrieve semantically similar image embeddings by returning images with high cosine similarity to it, and vice versa. The VGS model's general performance is then evaluated with a caption-to-image Recall@N score for image-caption pairs, representing the percentage of captions for which the correct image was found in the top N retrieved images [3].

Harwath and colleagues [3, 4] were among the first to propose such a neural model for the purpose of embedding images and speech, by training their network on the Flickr8k dataset which contains image-caption pairs. A few years later, a number of researchers, such as Chrupała and colleagues, trained VGS models on the same dataset. They improved the VGS model's architecture for embedding image and speech pairs, and achieved better caption-to-image retrieval scores in comparison with [3, 4] (R@1: 5.5). Merkx and

colleagues followed a few years later with an improved architecture and achieved a considerably better performance on the caption-to-image retrieval task (R@1: 8.0) [6]. Currently, Ilharco and colleagues have achieved the best caption-to-image retrieval scores by pretraining their VGS model on synthetic data and finetuning their model on the Flickr8k image-caption data (R@1: 13.9) [10].

A variety of different VGS architectures and tasks has been proposed since then. For example, Kamper et al. [7] introduced the task of semantic keyword spotting through VGS models. Based on a text query, these models aimed to retrieve semantically relevant speech utterances [66]. This task was later extended in order to perform in a cross-lingual setting, where Kamper et al. showed how the VGS model could retrieve a speech utterance in English from a German query [9]. Furthermore, some researchers have directed their focus on the representations and activations learned within VGS models and have discussed the methodology for their analysis [12–14, 67].

In the following section the neural networks used in the dual-encoder structure of a visually grounded speech model will be discussed. Afterwards, a number of ways in which word recognition with Visually Grounded Speech models has been researched will be discussed.

2.3.1. VISUALLY GROUNDED SPEECH MODEL ARCHITECTURES

As shown in Figure 2.6, the architecture of a VGS model consists of two encoders, one to encode speech data and another to encode image data. This dual-encoder setup allows such a model to encode two different modalities, such as images and speech [3–9], and train the model in order for it to embed the image-caption pairs together with high cosine similarity. First, a number of image encoders used in VGS models will be discussed, followed by a number of speech encoders.

IMAGE ENCODERS

The purpose of an image encoder is to create an image embedding which captures meaningful visual constituents within an image, such as the objects present within an image [3]. As a result, a speech signal can be aligned to these visual constituents, and the image and speech encoders can be trained simultaneously in order to create a visual-semantic alignment between images and speech. In order to find these visual constituents and create these image embeddings, often DNNs are applied which have been developed for large-scale image recognition tasks [3, 5–7], and can therefore recognise a large number of different classes, i.e. categories of objects, within images. Often, these image recognition networks are pre-trained on the Imagenet dataset [3–9], which contains 1.2 million images from 1000 different classes [68]. These pre-trained image recognition networks are then employed as image encoders by removing the final classification layer, and taking the neuron activations in the penultimate hidden layer as a representation of the visual constituents in an image.

In the VGS model by [3], the authors employed a regions convolution neural network (R-CNN) pre-trained on the Imagenet dataset [68, 69]. This R-CNN combines a region proposal module together with a traditional convolution neural network which looks for objects within the proposed regions. At the time, this network achieved top-performing scores on the Pascal VOC2012 object classification challenge, making it suitable for en-

coding images [69]. In later VGS architectures, researchers often opted for the VGG-16 image recognition network as their image encoder [5, 7, 12, 70]. VGG-16 is a deep CNN which shows outstanding performance on large-scale image recognition tasks, achieving top scores on the ILSVRC-2014 challenge for classifying images in Imagenet [71]. This architecture made improvements over previous image recognition networks by applying a very deep architecture with many layers, using smaller kernels for convolution [71]. In more recent VGS models, researchers have been using residual neural networks (ResNet) for their image encoders [6, 10], which have helped in achieving the best caption-to-image Recall@N scores in VGS models. Resnet also makes use of the improvements made with VGG-16, namely smaller kernels in the convolution layers, and a deep architecture [72]. Resnet employed an architecture eight times deeper than VGG-16, while having a lower training complexity, by employing residual connections between layers [72]. As mentioned in Section 2.2.2, the vanishing gradient problem can prevent deep architectures from learning during backpropagation [51]. The residual connections within ResNet allow the network to jump over hidden layers, lowering the complexity within the model, which helps the model to overcome this vanishing gradient problem during training. In [6], the authors employ a Resnet network with 152 layers pre-trained on Imagenet [72], with the final classification layer removed. Out of the networks pre-trained on Imagenet, which are readily available in popular Python Deep Learning frameworks (Pytorch, Keras), Resnet-152 has one of the lowest image classification error rates [72].

SPEECH ENCODERS

A speech encoder can be used to create speech embeddings of acoustic features. The purpose of the encoder is that it can find relevant constituents in a speech signal, such as words, after having been trained on image-caption pairs. Given that speech is inherently temporal, researchers often opt for DNN architectures, such as CNNs or RNNs, which model this type of information well [3, 5–7, 12]. Similar to the image encoder, the speech encoder should output a high-level representation of information in the speech signal, such as a high-dimensional vector of neuron activations. In most cases, the speech encoder is not pre-trained for speech classification, and the entire learning process happens during training of the VGS model. In the following paragraph, a number of architectures will be discussed that are used for this purpose.

In the work by Harwath and colleagues [3], a Convolutional Neural Network was used. Their CNN performed convolution and pooling operations over an acoustic feature, log mel filterbank spectrograms, which were made from the speech caption data. As explained in Section 2.2.2, using a CNN allows the VGS model to capture speech information from pre-segmented spoken words with some degree of invariance to speaker and environment. Harwath and colleagues trained their VGS model by using pre-segmented spoken words, as opposed to the speech captions often trained on in works that followed in later years [5–7, 12]. With some adjustments, the authors of [3] also trained their speech encoder using speech captions [4], which is a more difficult task since the speech captions are unsegmented spoken sentences, containing considerably more acoustic information about the visual constituents present in an image.

As more speech information was now present within the speech caption data, researchers started to consider speech encoder architectures which could better capture long-range time dependencies in speech signals. Chrupała and colleagues [5] made

some improvements to the dual-encoder setup by Harwath and colleagues [3] by changing the speech encoder from a CNN to a Recurrent Highway Network (RHN) [73]. Recurrent Highway Networks are a generalization of RNNs with gated recurrent units, that allow recurrent connections between neurons to be of multiple steps [73]. This RHN setup substantially improved Recall@N caption-to-image retrieval scores over the CNN setup in [3], most likely due to RHNs being able to more successfully model the temporal nature of speech signals [5]. After the work by Chrupała and colleagues using RHNs [5], a substantial improvement to the performance of VGS models was made by Merkx and colleagues by using a speech encoder which combines convolution operations together with an RNN [6]. This particular setup performs a convolution operation over an acoustic feature, by passing on the convolution frames to a number of stacked GRU units, after which the frames are passed through an attention mechanism in order to create the final speech embedding [6]. Combining these convolution, GRU and attention mechanisms resulted in an improvement over previous VGS architectures, achieving substantial improvements to Recall@N scores in the caption-to-image retrieval task (R@1: 8.0).

2.3.2. WORD RECOGNITION WITH VISUALLY GROUNDED SPEECH MODELS

In recent years, the question has arisen whether VGS models implicitly learn words from the speech captions and images they were trained on. [5, 6, 12–14]. For example, some research has shown that VGS models implicitly learn to recognise meaningful sentence constituents such as phonemes and words from the full length speech captions which they were trained on [5, 6, 12–14, 67]. This gave an indication that, although VGS models were trained on speech captions without supervision to recognise words, they could still potentially be used for word recognition.

In order to further understand how ASR systems work, as proposed by [17], it could be interesting to apply analyses and insights from HSR to develop research methodologies. In a paper by Havard and colleagues, inspiration was drawn from HSR research to investigate word recognition in a VGS model [12]. Using a dataset of synthetically spoken captions paired with images, they trained a VGS model in order to embed image-caption pairs with high cosine similarity. They investigated word recognition by presenting the model with isolated words and counted a correct word recognition if the model retrieved an image containing the correct visual referent [12]. They found that their model, which was trained on synthetically spoken captions and image pairs, correctly recognised words 8 out of 10 times when the model was presented with an isolated word, and was 'asked' to retrieve 10 images [12]. This showed that the model, which was trained on synthetic speech, did not just learn to encode these constituents, but the model actually 'recognised' individual words and learned to map them onto their correct visual referents. Furthermore, through a gating experiment where words were presented to the VGS model in increasing length from word onset and from word offset, they found that the model needed access to the first phoneme of a word in order to recognise a word [12]. This indicated that word recognition in their VGS model functioned similar to how word recognition is theorized to work in humans according to the COHORT model (Section 2.1.1).

3

METHODOLOGY

3

THE purpose of this research was to establish whether and how a Visually Grounded Speech model can perform word recognition using naturally spoken speech. This chapter describes the methodology used to answer the research questions defined in Section 1.1.

To answer the first research question, Experiment 1 (Section 3.3.2) evaluates whether a VGS model can perform word recognition using naturally spoken speech. In order to find out whether the VGS model can recognise individual words, individual word embeddings and image embeddings were created using the VGS model (Section 3.1). Afterwards, word recognition was evaluated through retrieving image embeddings with a high cosine similarity to the individual word embeddings, and using the words in the textual captions of the images as ground-truth labels.

To answer the second research question, Experiment 2 (Section 3.3.3) investigated how words are recognised in the VGS model. The VGS model was presented with words in speech segments of increasing duration i.e. words with an increasing number of phonemes. Each segment of phonemes was embedded and word recognition was evaluated through retrieving image embeddings with a high cosine similarity to the embedding of the segment of phonemes. This was done to gain insight into 1) the time-course of word recognition, 2) the amount of information needed for word recognition, and 3) whether the neural model is able to encode units of sound, such as phonemes.

Furthermore, to answer the second research question, Experiment 3 (Section 3.3.4) was performed to research how word recognition in the VGS model is affected by providing a preceding context to the word. This third experiment was inspired by priming experiments done in HSR research. In this experiment a contextual stimulus, i.e. the caption preceding the word, was encoded in the VGS model before word recognition was evaluated. This was done to gain insight into how contextual information affects word recognition in a VGS model.

Lastly, in order to better understand linguistic and acoustic factors affecting word recognition in a VGS model, a statistical analysis of factors affecting word recognition was done for the results of the first and second experiment (Section 3.3.2 & Section 3.3.3).

In the next three chapters, firstly the VGS model with which the three experiments were performed will be illustrated. Secondly, the metrics used to evaluate the model's caption-to-image retrieval performance and word recognition performance will be discussed. Lastly, the three experiments using this VGS model will be described.

3.1. MODEL CONFIGURATION

For the evaluation of word recognition, a version of the RNN-based VGS model configuration by Merkx and colleagues was used [6]. The architecture of the model used is shown in Figure 3.1. The model consists of a dual-encoder structure of two Deep Neural Networks (DNNs): an image encoder and a caption encoder. The encoders embed the speech captions and images and the model is then trained to minimise the cosine distance between image-caption pairs in the embedding space.

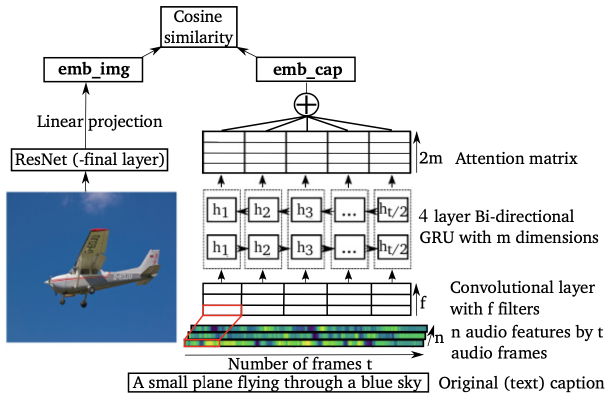


Figure 3.1: A visual representation of the image encoder parallel to the caption encoder. Adapted from [6].

3.1.1. ACOUSTIC FEATURES

As explained in Section 2.2.3, transforming a speech caption to an acoustic feature allows the representation of acoustic properties of a speech signal in a way that can be given as input to a DNN. For this research MFCCs were chosen as the acoustic feature used to train and test the model. MFCCs have shown great performance in speech recognition tasks, which is partly attributed to MFCCs approximating the non-linearity of the human auditory system [59]. The manner in which MFCCs are computed is explained in Section 2.2.3. Using code developed by [6], these features were computed for the speech captions in the Flickr8k Dataset.

The MFCCs were calculated over 25ms frames, with a shift of 10ms. Of the 14 MFCC features computed, the first MFCC feature was discarded for lack of conveying relevant information to the shape of the spectrum, in order to arrive at a set of 13 MFCC features. The first-and second-order variations of these MFCC features were concatenated to these features, resulting in 39-dimensional feature vectors representing each frame in a speech caption.

3.1.2. IMAGE ENCODER

The image encoder used in this VGS model makes use of ResNet-152, which was pretrained on ImageNet, a dataset containing a 1000 classes and 1.2 million images [72]. As discussed in Section 2.3.1, ResNet-152 is a deep convolutional neural network with out-

standing performance on image classification tasks [72], which makes it suitable as an image encoder.

ResNet-152 expects all input images to be of the same dimensions. For that reason, all images in Flickr8k were first standardized to the same size. In order to retain aspect ratio, the images were resized so that the smallest side contained 256 pixels. As some images did not have equal dimensions, some visual information could be lost if the image would be cropped. To avoid this, 5 different 224 x 224 pixel crops were taken from all four corners and from the middle of the image. Furthermore, the same 5 crops were taken from the mirrored image to create more data points.

The resulting 10 crops were passed through ResNet-152. In Figure 3.1 it is shown that the final classification layer of ResNet-152 was removed and the vectors of activations in the penultimate layer was taken. These feature vectors were computed for all 10 crops of an image and were then averaged out.

As can be seen in Figure 3.1, a single layer linear projection was consecutively performed on top of the averaged out feature vector from ResNet-152. This layer of the image encoder could then be trained to minimize the loss between the training set images and captions. As a result, the ResNet-152 architecture together with this linear projection could be used to create image embeddings.

3.1.3. CAPTION ENCODER

The caption encoder used in this research consists of three main parts: a convolutional layer, 4 layers of GRUs and an attention layer. This caption encoder is an adaptation of the configuration used by Merkx and colleagues [6]. For this research, an additional GRU layer was added to their configuration, which further deepens the architecture of the caption encoder.

As input, the caption encoder receives MFCC features generated for the speech captions in the Flickr8k training dataset. Firstly, a 1-dimensional convolutional layer was applied to the MFCC features with a kernel of size 6 and stride of 2 and 64 output channels. These channels were then fed to a 4-layer bi-directional GRU, which, as opposed to a uni-directional GRU (Section 2.2.2), allows the model to learn long-range time dependencies from left-to-right and right-to-left in the MFCC features [6]. Afterwards, the 1024 bi-directional GRU units were concatenated to create 2048-dimensional feature vectors, which were fed into the attention layer which computed a weighted sum for all captions. Finally, the resulting feature vectors were L2 normalised to arrive at the final caption embeddings.

3.1.4. LOSS FUNCTION AND HYPER-PARAMETERS

The dual-encoder structure of the VGS model allows a loss to be calculated between image-caption pairs passed through the image and caption encoders. As in the work by Merkx and colleagues, a hinge loss function was adopted (see Equation 3.1) as a guidance for backpropagation [6].

$$l(\theta) = \sum_{(c,i),(c',i') \in B} (\max(0, cos(c,i') - cos(c,i) + \alpha) + \max(0, cos(i,c') - cos(i,c) + \alpha)) \tag{3.1}$$

$l(\theta)$ represents the loss calculated over the network parameters θ for a mini-batch B. The model was given mini-batches B of 32 correct caption-image pairs (c,i), where for each correct pair (c,i) the other caption-image pairs in the batch were used to make mismatched pairs (c,i') and (c',i) [6]. The mismatched pairs were made by taking the 25% 'hardest', i.e. highest cosine similarity, mismatched pairs (c,i') and (c',i) for each caption-image pair (c,i). The VGS model was trained to embed caption-image pairs to have a cosine similarity larger by a margin $\alpha = 0.2$ than the cosine similarity for mismatched pairs.

Lastly, Adam [74] was used as the optimization algorithm, and a cyclic-learning rate [75] was used which moves smoothly between a minimum and maximum bound of 10^{-6} and $2 * 10^{-4}$. The model was trained for 32 epochs.

3.1.5. DATASET

The model was trained on the Flickr8k dataset. This dataset contains 8000 images from the Flickr photo-sharing website, which have been paired with 5 written captions per image [76]. The resulting dataset contains 8000 images with 40,000 textual captions depicting a wide variety of everyday actions and events [76]. Spoken versions of these captions were collected from 183 different speakers in a crowdsourcing effort by Harwath and colleagues [3]. For the training, validation and test set, the data split provided by [65] was used. This split the dataset into a training set of 6000 images, a validation set of 1000 images and a test set of 1000 images.

3.2. EVALUATION METRICS

3.2.1. EVALUATING CAPTION-TO-IMAGE RETRIEVAL

After training, the general performance of VGS models is often evaluated by calculating a caption-to-image Recall@N score [3–10], which represents how well the model returns a matching image given a speech caption. As explained in Section 2.3.1, Recall@N (R@N) represents the percentage of speech captions for which the correct image was in the top N highest cosine similarity images returned. Furthermore, the Median R is used, which is the median rank of the correctly retrieved image given a speech caption.

3.2.2. EVALUATING WORD RECOGNITION

In the paper by Havard and colleagues [12], the retrieval of an image containing a word's correct visual referent was used as a measure of the model's word recognition performance. In order to quantify this, a Precision@10 (P@10) score was calculated for each tested word or for the tested word's phoneme sequences. P@10 was used as opposed to R@10, because multiple correct images can be returned containing a correct visual referent given a tested word.

P@10 was calculated by returning the 10 image embeddings with the highest cosine similarity to the embeddings of the tested word or for the tested word's phoneme sequences. P@10 was then calculated as the percentage of the 10 images which contained the correct visual referent. As Flickr8k does not contain labels for image content, the words in the image captions were used as ground-truth labels.

3.3. EXPERIMENTS

In this section, the three experiments, which were performed to investigate word recognition and answer the research questions, will be outlined. To perform the experiments, four speech corpora were created. Firstly, it will be discussed how these datasets were created. Secondly, the three experiments using these datasets will be described.

The code, used to create the datasets and perform the experiments, was written in Python 3.7 and R. The code of the VGS model used in the experiments can be found on this Github page [1]. The code used to create the speech corpora, the Python classes to interact with the VGS model, and the code used to perform the experiments and analyse the results can be found on this Github page [2].

Table 3.1: The words that are tested for word recognition in the model, together with the phonemes the word comprises of

	Word	Phoneme Sequence		Word	Phoneme Sequence
1	dog	['D', 'AO', 'G']	26	hair	['HH', 'EY', 'R']
2	man	['M', 'AE', 'N']	27	football	['F', 'UH', 'T', 'B', 'AO', 'L']
3	boy	['B', 'OY']	28	sunglasses	['S', 'AH', 'N', 'G', 'L', 'AE', 'S', 'AX', 'Z']
4	girl	['G', 'ER', 'L']	29	head	['HH', 'EH', 'D']
5	snow	['S', 'N', 'OW']	30	shorts	['SH', 'AO', 'R', 'T', 'S']
6	people	['P', 'IY', 'P', 'XL']	31	basketball	['B', 'AE', 'S', 'K', 'IH', 'T', 'B', 'AO', 'L']
7	dogs	['D', 'AO', 'G', 'Z']	32	table	['T', 'EY', 'B', 'XL']
8	shirt	['SH', 'ER', 'T']	33	water	['W', 'AO', 'T', 'AXR']
9	child	['CH', 'AY', 'L', 'D']	34	grass	['G', 'R', 'AE', 'S']
10	ball	['B', 'AO', 'L']	35	bench	['B', 'EH', 'N', 'CH']
11	person	['P', 'ER', 'S', 'AX', 'N']	36	woman	['W', 'UH', 'M', 'AX', 'N']
12	pool	['P', 'UW', 'L']	37	air	['EY', 'R']
13	men	['M', 'EH', 'N']	38	field	['F', 'IY', 'L', 'D']
14	girls	['G', 'ER', 'L', 'Z']	39	street	['S', 'T', 'R', 'IY', 'T']
15	bike	['B', 'AY', 'K']	40	mouth	['M', 'AW', 'TH']
16	rock	['R', 'AA', 'K']	41	dirt	['D', 'ER', 'T']
17	face	['F', 'EY', 'S']	42	mountain	['M', 'AW', 'N', 'T', 'XN']
18	boys	['B', 'OY', 'Z']	43	children	['CH', 'IH', 'L', 'D', 'R', 'AX', 'N']
19	hat	['HH', 'AE', 'T']	44	ocean	['OW', 'SH', 'AX', 'N']
20	player	['P', 'L', 'EY', 'AXR']	45	sand	['S', 'AE', 'N', 'D']
21	jacket	['JH', 'AE', 'K', 'IH', 'T']	46	building	['B', 'IH', 'L', 'D', 'IX', 'NG']
22	dress	['D', 'R', 'EH', 'S']	47	soccer	['S', 'AA', 'K', 'AXR']
23	swing	['S', 'W', 'IH', 'NG']	48	park	['P', 'AA', 'R', 'K']
24	car	['K', 'AA', 'R']	49	camera	['K', 'AE', 'M', 'AXR', 'AX']
25	wall	['W', 'AO', 'L']			

[1] https://github.com/DannyMerkx/speech2image
[2] https://github.com/sebastiaansch/WordRecognitionInVGSmodel

3.3.1. SPEECH CORPORA

The speech corpora used for the experiments have been created from the 5000 speech captions and 1000 images in the Flickr8k test set. To perform the three word recognition experiments, four speech corpora were needed: a dataset of words, two datasets of phoneme sequences of those words and a dataset containing the preceding captions to the tested words. In the following sections, the selection criteria for the tested words will be explained, followed by a description of how the datasets were created.

TESTED WORDS

A visually grounded model relies on there being a consistency between the image content and speech content in order to create a common embedding space. Therefore, a set of 49 nouns (Table 3.1) which contained clear visual referents was chosen, such as 'bike' and 'man', as opposed to articles and adverbs. Furthermore, in order to ensure each word had enough accompanying images, each noun had to occur at least 50 times in the test set captions. For this research, the selected 49 words will be referred to as 'word types'.

SEGMENTING TEST DATA

Of each word type, the first 50 occurrences were segmented from the speech captions in the test set. In this research, an occurrence of a word type will be referred to as a 'word token'.

To segment the word tokens, the word tokens' phoneme sequences, and the captions preceding the word tokens, a forced alignment of the speech captions in the Flickr8k test set was used [3]. This forced alignment contained phoneme boundaries of the words in the speech captions of the Flickr8k test set.

As a result, the first dataset of word tokens was created by segmenting the speech captions at their word boundaries. The second dataset was created by segmenting each word token at its phoneme boundaries at increasing lengths from word onset. For example, for the word 'bike', the speech signal was segmented into 'B', 'B-AY', and 'B-AY-K'. The third dataset was created by segmenting each word token at its phoneme boundaries at increasing lengths from word offset. For example, for the word 'bike', the speech signal was segmented into 'K', 'AY-K', and 'B-AY-K'. The fourth dataset was created by segmenting the part of the caption that preceded each word token.

[3] The forced alignment was created by Markus Müller (https://scholar.google.de/citations?user=t1g04OYAAAAJ&hl=en). Permission was granted for usage.

3.3.2. EXPERIMENT 1: WORD RECOGNITION

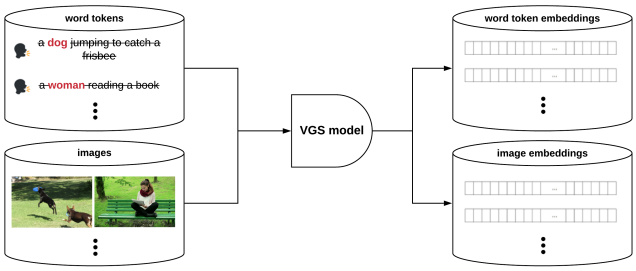


Figure 3.2: A visual representation of how the word token dataset is used in the first experiment

The aim of the first experiment was to determine whether the VGS model can recognise words. In order to do this, the 49 word types shown in Table 3.1 were tested. For this experiment, the first dataset containing the 50 word tokens for each word type was used. Furthermore, images were taken from the Flickr8k test set to test word recognition on, namely the images from which word tokens were segmented from its speech captions.

As shown in Figure 3.2, the word tokens and images were passed through the trained VGS model in order to create embeddings of both the word tokens and images. For each word token embedding, the ten image embeddings with the highest cosine similarity to the word token embedding were taken, and a P@10 score was calculated to represent how many images contained the correct visual referent.

To examine linguistic and acoustic factors which influence the model's word recognition performance, a Linear Mixed Effects Regression (LMER) was performed. An LMER is a suitable regression model to answer this question because it allows the modelling of random effects. This allows the modelling of variability caused by different speakers and different words, through the calculation of different intercepts for each random effect. For the LMER analysis the *lme4* package in R was used [77]. All fixed effects that were tested have been Z-Score normalised. The dependent variable tested is the P@10 score.

For the word recognition experiment, the LMER model takes in as fixed effects: the signal duration (i.e., number of speech frames), the speaking rate calculated as the number of phonemes in the word divided by its signal duration, the frequency of occurrence of the word in the training set and the number of phonemes, vowels, and consonants in the word. The linguistic and acoustic features used as fixed effects are known to influence human speech processing (Section 2.1.2). The *lme4* formula used to test the factors affecting word recognition is shown in Figure 3.3 (For more information on how this represents the mathematical description of the LMER model please refer to [78]).

```
P@10 ~ Signal duration + Speaking rate + Training set word frequency *
(\# of Vowels + \# of Phonemes + \# of Consonants) + (1 | Speaker ID)
+ (1 | Word ID) + (0 + Signal Duration | Speaker ID)
```

Figure 3.3: The *lme4* formula used to investigate factors affecting word recognition

The two-way interaction of the frequency of occurrence of the word in the training set with the number of phonemes, vowels, and consonants was also included. These interaction effects were taken into consideration because words with a certain number of phonemes, vowels, and consonants might appear more often than others in this dataset. Furthermore, by-speaker and by-word random intercepts as well as by-speaker random slopes for signal length were included so that speaker differences with regard to the duration of the signal would also be taken into consideration.

3.3.3. EXPERIMENT 2: TIME-COURSE OF WORD RECOGNITION

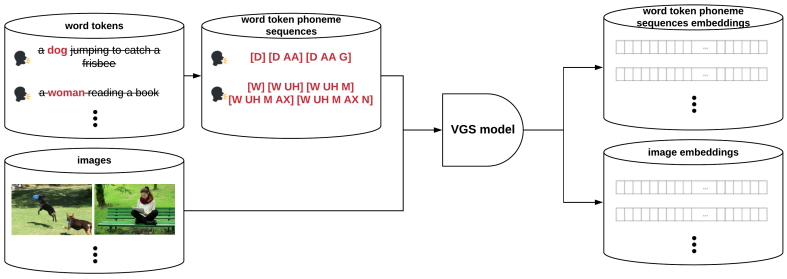


Figure 3.4: A visual representation of how the dataset of word tokens phoneme sequences is used in the second experiment

The aim for the second experiment was to investigate how words are recognised in the VGS model. A gating paradigm, borrowed from human speech processing (Section 2.1.3), was used for that purpose. Words were presented to the VGS model in speech segments of increasing duration, i.e., with an increasing number of phonemes, and the model was 'asked' to retrieve an image of the correct visual referent on the basis of the available phoneme string. This was done to reveal 1) the time-course of word recognition, 2) the amount of information needed for word recognition, and 3) whether the neural model is able to encode units of sound, such as phonemes.

For this experiment, the second dataset, which contained words tokens segmented at their phoneme boundaries, was used. For each word token, this dataset contains phoneme boundaries segmented at increasing lengths from word onset. Furthermore, the images from the Flickr8k test set from which the phoneme sequences were segmented from its speech captions were taken.

As shown in Figure 3.4, the word tokens segmented at their phoneme boundaries from word onset were passed through the VGS model together with the images, in order to create embeddings of both. For each word token embedding (at increasing length), the ten image embeddings with the highest cosine similarity to it were taken, and a P@10 score was calculated to represent how many images contained the correct visual referent.

In order to determine the difference in word recognition in the VGS model from word onset in comparison to word offset, the process as displayed in Figure 3.4 was also applied to the third dataset of word tokens segmented at their phoneme boundaries from word offset. This was done to compare whether words are recognised better from word

onset or word offset. As a result, the time-course of word recognition could be compared with regard to words presented from word onset and words presented from word offset.

This experiment makes use of an LMER to evaluate the linguistic and acoustic factors affecting word recognition. The LMER model takes into account the earlier mentioned frequency of occurrence of the word in the training set and the total number of phonemes in the word. The size of the word-initial cohort and neighbourhood density were also included. The word-initial cohort was calculated by determining the number of words for each phoneme sequence that started with the same phoneme sequence in the Flickr8k training set, which contains a total of 6182 unique words. This factor indicates the number of words that are considered simultaneously for recognition by the model given the phoneme sequence it has seen so far. The neighbourhood density is calculated as the number of words - from the words in the Flickr8k training set - that can be formed from the phoneme sequence by a one-phoneme substitution [79]. This factor indicates the similarity between spoken forms of words, and is therefore a second measure of the number of words that are simultaneously considered for recognition. As in the previous experiment, these factors used as fixed effects are known to influence human speech processing (Section 2.1.2). The model also includes a by-speaker and a by-word random intercept. The *lme4* formula used to test factors affecting the time-course of word recognition is shown in Figure 3.5

```
P@10 ~ \# of Phonemes + Word-initial cohort + Neighbourhood density +
Training set word frequency + (1 | Speaker ID) + (1 | Word ID)
```

Figure 3.5: The *lme4* formula used to investigate factors affecting the time-course of word recognition

3.3.4. EXPERIMENT 3: WORD RECOGNITION WITH PRECEDING CONTEXT

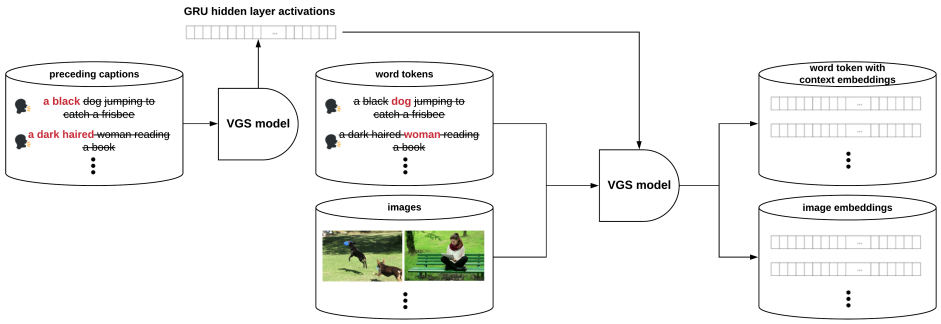


Figure 3.6: A visual representation of how the fourth dataset of preceding captions is used to 'prime' the GRU in the VGS model used to evaluate word recognition.

The third experiment investigates how word recognition is affected by preceding context. This experiment is inspired by priming experiments performed in human speech processing research (Section 2.1.3). In a priming experiment, the goal is to see whether

exposure to one stimulus influences a response to a subsequent stimulus. This experiment can reveal if, like in humans [17], contextual information can aid or hinder word recognition in a VGS model.

The word tokens of the first dataset have been tested on word recognition in both a 'primed' and 'unprimed' setting. In the first setting the model was exposed to the preceding caption and in the second setting it was not given any contextual information.

To provide contextual information, the fourth dataset of captions preceding the word tokens has been used. As shown in Figure 3.6, in the 'primed' setting, the VGS model has been passed the captions preceding the word tokens, and the hidden activations from the GRU are taken. As shown in Figure 3.6, these hidden states are then used as initial hidden states for the GRU in the VGS model which is used to embed the word tokens. As a result, the word tokens were essentially 'primed' with the contextual information from the captions which preceded them. In the 'unprimed' setting, the word tokens were simply passed through the VGS model without an initial hidden state in the GRU from the preceding caption. As in the previous experiments, Experiment 3 uses the images from the Flickr8k test set from which the preceding captions and word tokens were segmented. For each 'primed' and 'unprimed' word token, the ten image embeddings with the highest cosine similarity to the word token embedding were taken, and a P@10 score was calculated to represent how many images contained the correct visual referent.

4

Results

THIS chapter begins with an evaluation of the general performance of the VGS model with regard to caption-to-image R@N scores. In order to put the R@N scores in perspective, the results are compared to caption-to-image R@N scores of other researchers' VGS models trained on the Flickr8k dataset. Subsequently, the results are organised and presented following the order of Experiments 1,2 and 3:

1. Word Recognition
2. Time-course of Word Recognition
3. Word Recognition with Preceding Context

4

4.1. CAPTION-TO-IMAGE RETRIEVAL

Table 4.1 shows the results for the caption-to-image retrieval task obtained with the model used in this research, as presented in [80]. For comparison, the results of some other well-known VGS models trained on Flickr8k are listed in the same table. The addition of an extra GRU layer has led to a substantial performance increase when compared to the original architecture made by [6], most likely because long-range dependencies could be captured better and because the architecture of the speech encoder was further deepened.

Table 4.1 shows that, in comparison with Merkx et al. [6], Chrupala et al. [5] and Harwath et al. [3], whose models had also only been trained on Flickr8k, this model achieves the highest R@N and Median R scores. Table 4.1 also shows that only Ilharco and colleagues achieved a higher R@N score than this model [10]. However, a notable difference from the other models shown in Table 4.1 is that Ilharco and colleagues' image and speech encoders had been pre-trained on a large dataset of synthetically spoken captions and images, which was later fine-tuned on the captions and images of Flickr8k.

Table 4.1: Caption-to-image retrieval scores including 95% confidence intervals for the model used in this research, as presented in the paper [80], as well as other well known models trained on the Flickr8k dataset.

Model	R@1	R@5	R@10	Med. R
Ilharco et al. [10]	13.9±2.3	36.8±3.1	49.5±3.2	-
Scholten et al. [80]	10.7±1.9	29.2±2.8	40.2±3.0	18
Merkx et al. [6]	8.0±1.7	24.5±2.7	35.5±3.0	24
Chrupała et al. [5]	5.5±1.4	16.3±2.3	25.3±2.7	48
Harwath et al. [3]			17.9±2.4	

4.2. EXPERIMENT 1: WORD RECOGNITION

In the word recognition experiment, for each word type, 50 word tokens were presented to the model. In the bar chart in Figure 4.1, the blue bar represents the average P@10 score for a word type and the yellow bar shows the highest P@10 score achieved for a word type. The average P@10 calculated over all word types is 0.44, which indicates that on average 4.4 out of the ten retrieved images contained the correct visual referent. Figure 4.1 also shows that a number of word types, such as 'face', 'head' and 'hair', have a P@10 near zero, which means that no correct images were retrieved and that the words were not recognised. Havard and colleagues [12] reported a median P@10 of 0.8, while in this research a median P@10 of 0.4 was achieved as the average word type score. If the best performing word is taken from each word type, a median P@10 of 0.8 is achieved with this model.

While this model does learn to recognise most words to some degree, the average word type P@10 scores indicate a considerable difference in recognition performance between the synthetically spoken words used in [12] to the naturally spoken words of Flickr8k.

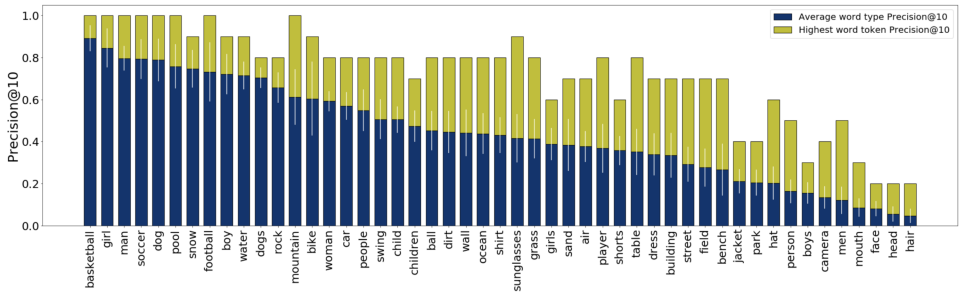


Figure 4.1: Average and highest P@10 scores for the 49 tested word types. The white line in the bar represents the standard deviation of the average P@10 score.

In order to gain an understanding of what some words might be 'mistaken' for by the model during the word recognition experiment, the cosine similarity between the word type embeddings averaged over its word tokens are shown in Figure 4.2. As shown in Figure 4.2, the embeddings of the word types are often close in cosine similarity to word types that are phonemically similar or where a part of a word sounds similar, such as for 'head' and 'hat', 'dirt' and 'shirt' or 'hair' and 'camera'. Also, words are close in cosine similarity to words they often share an image with and thus likely share some semantic relatedness with, such as for 'park' where its word type embedding is close in cosine similarity to the word type embeddings of 'person', 'hat', 'shirt', 'water', 'air'. This suggests that words are embedded in a manner that represents both the phonemes that the word consists of, and a semantic closeness to words it often shares a caption or image with.

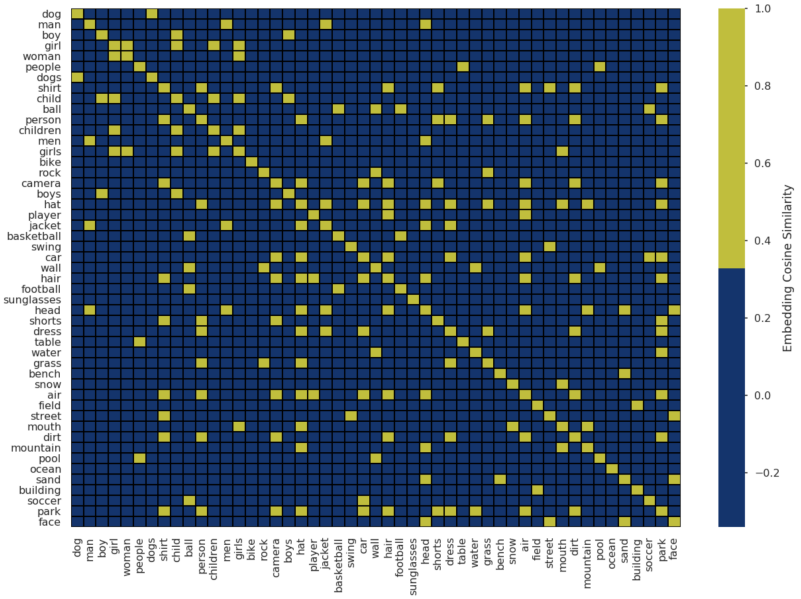


Figure 4.2: Cosine similarity between the word type embeddings averaged over its word tokens

Table 4.2: Significant fixed effects with Standard Errors for the word recognition LMER.

Fixed effects	Estimate	P-value
Intercept	0.432±0.033	<0.001
Signal duration	-0.050±0.014	<0.001
Speaking rate	-0.068±0.013	<0.001
Training set frequency	0.152±0.063	0.020

Table 4.2 shows the significant fixed effects found with the LMER. Firstly, signal duration was found to have a significant negative effect on the P@10 scores. This indicates that the model has more difficulty recognising longer words. Secondly, speaking rate also had a significant negative effect, showing that words that are spoken more rapidly were recognised less well than words that were pronounced more slowly. Lastly, the frequency of occurrence of the word in the training set was found to have a significant positive effect on word recognition performance. This shows that words which occur more often in the training dataset are recognised considerably better. No main effects were found for the number of vowels, phonemes or consonants in a word. Also, no interaction effects were found between the training set frequency of a word and the number of vowels, phonemes or consonants in a word.

With regard to random effects, it transpires that the standard deviation in the scores caused by testing different words (0.19) is larger than the standard deviation caused by different speakers (0.06). This is as expected, given that variation caused by different samples of words is likely larger than between different speakers saying the same words.

4.3. EXPERIMENT 2: TIME-COURSE OF WORD RECOGNITION

In order to investigate the time-course of word recognition of the VGS model, phoneme sequences of increasing length were given to the model. Figure 4.3 shows the results in terms of the P@10 of a given word type (shown on the y-axis) as a function of percentage of phonemes of the word type shown from word onset (shown on the x-axis). Note that the x-axis has ten values. If a word has for instance only two phonemes, the P@10 for the first and second phoneme span 10-50% and 60-100% respectively. A more yellow colour corresponds to a higher P@10.

As shown in Figure 4.3, generally, the more phonemes of a word the model is exposed to, the better it can retrieve the image corresponding to the spoken word. The steepest increase in word recognition scores happens after the model is exposed to 30-40% of a word's phonemes, whereafter most words are recognised accurately. Some words are not recognised at all, irrespective of the percentage of phonemes shown to the model, which can be seen at the bottom of 4.3 where the bars are entirely blue. For some words, such as 'person' or 'men', word recognition was highest after the first phoneme, and decreased after more phonemes were presented to the model.

Table 4.3: Significant fixed effects with Standard Errors for the time-course of word recognition LMER.

Fixed effects	Estimate	P-value
Intercept	0.295±0.020	<0.001
# of phonemes	0.134±0.003	<0.001
Training set frequency	0.087±0.018	<0.001
Word-initial cohort	-0.037±0.003	<0.001

The significant fixed effects of the LMER are summarised in Table 4.3. Unsurprisingly, the number of phonemes in a phoneme sequence has a significant positive effect on the P@10 scores, indicating that words are recognised better when the model is presented with longer phoneme sequences. The frequency of occurrence of the word in the training set again has a significant positive effect on the performance, showing that having more training examples allows phoneme sequences to be mapped better to the correct visual referent. The size of the word-initial cohort has a significant negative effect on the P@10 scores, indicating that word recognition is more difficult when there are more words which share a phoneme sequence at the start of the word. The effect of the neighbourhood density was not found to have a significant effect on word recognition.

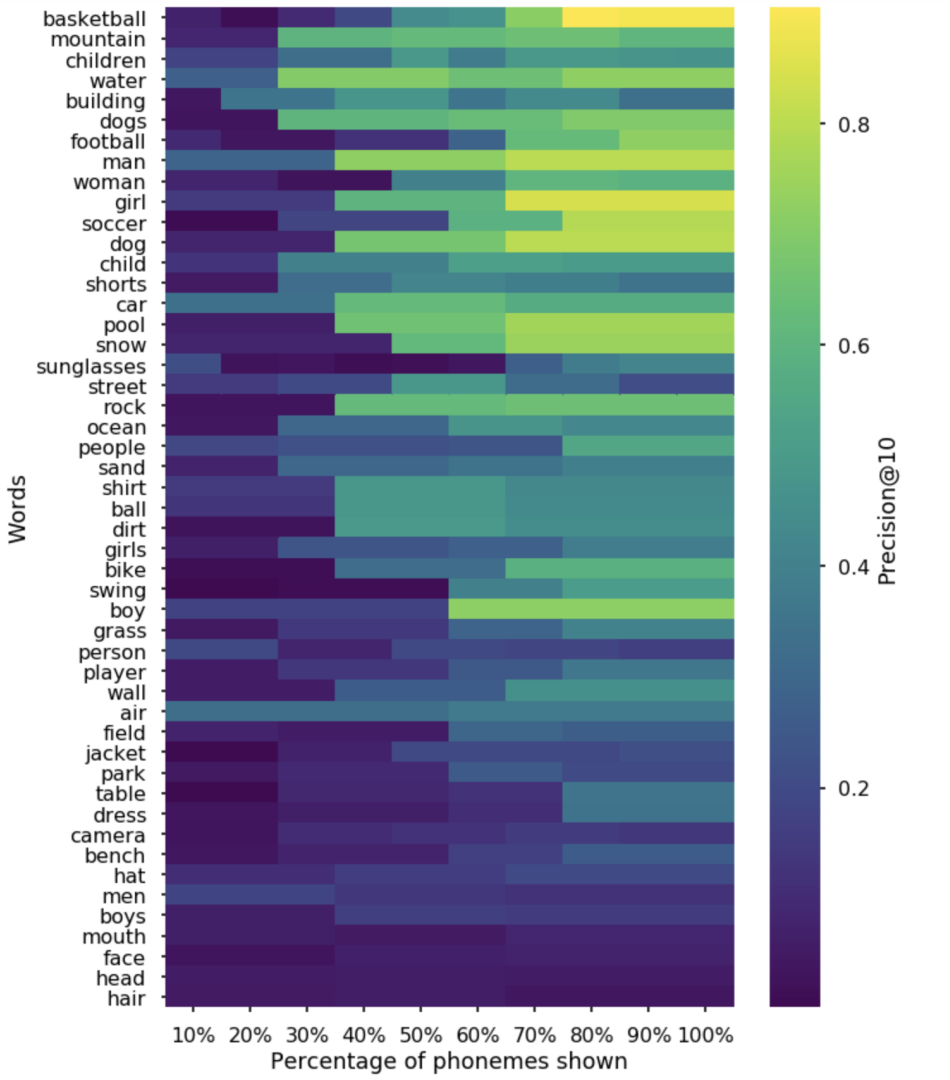


Figure 4.3: Heatmap of P@10 scores for a word type (shown on the y-axis) as a function of the phoneme sequence length, with the x-axis showing the percentage of phonemes of the word available to the model.

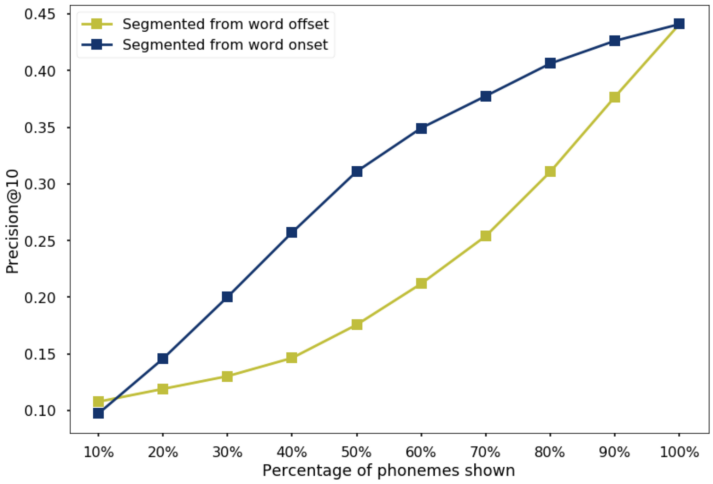


Figure 4.4: The lines represent the average P@10 score over all word types as a function of the percentage of phonemes shown either from word onset (blue) or from word offset (yellow).

Figure 4.4 shows the development of average P@10 scores for the percentage of phonemes shown to the model from word onset and word offset. The blue line represents the average P@10 score over all word types as a function of the percentage of phonemes shown from *word onset*, for example 'B', 'B-AY' and 'B-AY-K', and the yellow line the average P@10 score over all word types as a function of the percentage of phonemes shown from *word offset*, for example 'K', 'AY-K' and 'B-AY-K'. When words are presented from word onset, word recognition generally improves the most while the first 50% of a word's phonemes are being presented to the model. When words are presented from word offset, word recognition improves considerably faster when the model is presented with 60-100% of a word's phonemes.

Furthermore the yellow line is consistently below the blue line until all phonemes of a word are presented to the model. This shows that words are recognised more easily when their phoneme sequences are presented from word onset than from word offset. However, it seems the model can recognise words from both the word onset and offset.

Also, in Figure 4.3 it was shown that most words were recognised after having been exposed to 30-40% of a word's phonemes when presented from word onset. As shown in Figure 4.4, to achieve similar word recognition scores from word offset, the model would have to be exposed to 60-70% of a word's phonemes.

4.4. EXPERIMENT 3: WORD RECOGNITION WITH PRECEDING CONTEXT

In order to investigate whether 'priming' the VGS model with preceding context could improve the P@10 scores for the word tokens, a preceding context experiment was performed. Target words were given to the model in two settings. First only the word token was presented, and then the word token was presented while the hidden layer of the GRU was 'primed' with the word token's preceding caption. Figure 4.5 shows that words types that are generally recognised well and that have a high P@10 score tend to benefit less from being 'primed' with the preceding captions. Words that are in the bottom 50th percentile with regard to word token P@10 scores benefit substantially more from being 'primed'. On average, 'priming' relatively increased the P@10 scores by 6% for all word types. For the bottom 50th percentile, 'priming' relatively increased the P@10 scores by 27%.

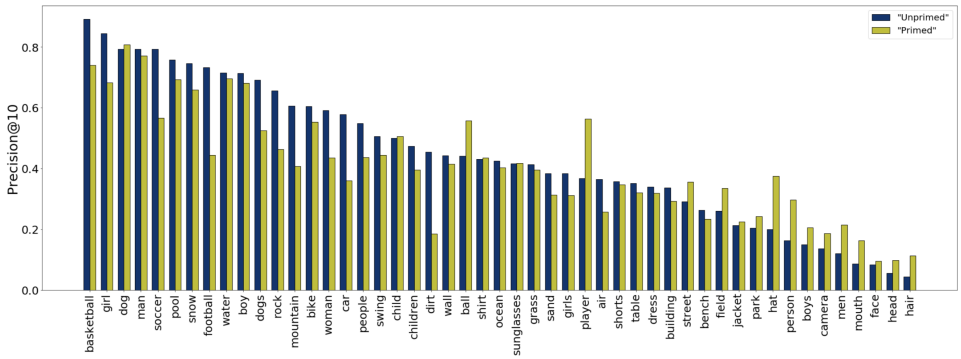


Figure 4.5: Average P@10 scores for each word type in the 'primed' and 'unprimed' setting.

5

Discussion and Limitations

5.1. DISCUSSION

THIS thesis investigated *whether* and *how* words are recognized by a Visually Grounded Speech model using real human speech. Word recognition was evaluated through retrieving image embeddings with a high cosine similarity to segmented word embeddings, while using the words in the textual captions of the images as ground-truth labels. To investigate word recognition in the VGS model, a methodology inspired by human speech recognition research was used for the design of three experiments: a word recognition experiment, a time-course of word recognition experiment and a word recognition with preceding context experiment.

The first experiment showed that, although the VGS model was trained on speech captions, it was able to encode information about individual words which it learned from these captions, and as a result, when given a word, could in most cases find images with correct visual referents. This suggests that during training time, the VGS model implicitly learned individual words from the speech captions, while not being explicitly supervised to do so. In this research, multiple word tokens of each word type were tested in order to compute an average word type recognition P@10 score, which resulted in a more accurate reflection of P@10 scores per word type for different speakers and pronunciations. As described in Section 2.3.2, Havard and colleagues performed their experiments using a synthetically generated caption dataset with different images than the ones used in this research (MSCOCO dataset [81]), and evaluated word recognition based on only a single instance of each word [12]. Synthetic speech contains less variation in quality, noise and speaking rate than naturally spoken speech. If the *average* P@10 scores for each word type attained in this research (0.44) is compared to the *single instance* P@10 score by Havard and colleagues (0.8), the score in this research is considerably lower. This suggests that the relatively 'optimistic' results shown in [12] do not accurately reflect the difficulty of the word recognition task, and that naturally spoken words pose a bigger challenge for word recognition than synthetically spoken captions.

There is one particular difference between the two datasets which might also have contributed to the difference in P@10 scores between [12] and this research. This research tested a set of 49 word types which were depicted in a wide variety of everyday actions and events in the training captions and images. Havard and colleagues tested 80 word types which were the main objects of each image and caption in the training dataset. Therefore, due to the larger variety of actions and events depicted in the dataset used in this research, it was likely more difficult for the VGS model to create a visual-semantic alignment between the images and captions.

Furthermore, the first experiment showed that average word type embeddings are often closest in cosine similarity to word types that are phonemically similar or where part of a word type sounds similar to another word type. Furthermore, it was found that word types were also embedded closely to word types they often shared an image with, and which as a result might share some semantic relatedness to it. This suggests that words are embedded in a manner that represents both the phonemes that the word comprises of, and a semantic closeness to words it often shares a caption or image with. This indicates that although the model is trained on caption-image pairs, it learns to capture both acoustic information and semantic meaning in its word embeddings.

The statistical analyses performed in the first and second experiment revealed a num-

ber of linguistic and acoustic factors affecting word recognition in VGS models. The most notable factors that were found to affect word recognition are: speaking rate, signal duration, training set word frequency and the size of the word-initial cohort. Some of these factors seem to affect word recognition in the VGS model in a similar manner to how they have shown to affect word recognition in humans (Section 2.1.2).

For example, during word recognition in humans, the more word competitors (Section 2.1) there are, the longer it generally takes for a word to be recognised. The size of the word-initial cohort was the factor used in this research to indicate the amount of word competition during word recognition, and this factor was found to negatively impact P@10 scores of a word. This indicates that the size of the word-initial cohort affects word recognition in VGS models similar to how it does in humans, by making word recognition more difficult.

It is also known that human recognise words that occur more frequently in their language faster and with lower error rates than words that occur less frequently in their language (Section 2.1.2). In this research the training set word frequency was used as a measure of how often a word had been seen by the VGS model when it was trained, which essentially displays how often it appears in the model's known vocabulary or 'language'. The word frequency was found to positively impact word recognition scores, showing similarities to how a word's frequency influences recognition speeds in humans.

This research found that speaking rate also affects word recognition in VGS models. Humans identify words less well if they are presented at mixed speaking rates than when they are presented at a single speaking rate (Section 2.1.2). In the VGS model, it was found that words that were spoken more rapidly were recognised less well. Although this result does not show a direct similarity to what has been found in behavioural research of human word recognition, it does show that speaking rate affects word recognition both in humans and VGS models.

The second experiment showed that the VGS model can recognise words after having been exposed to only a partial sequence of a word's phonemes. This experiment showed that the steepest increase in word recognition occurred after the model was presented with 30-40% of a word's phonemes from word onset. To achieve a similar word recognition performance from word offset, the model needs to have access to 60-70% of a word's phonemes. Similar to human listeners [18], the model did not need to have all phonemes of a word available in order to recognise it, which indicates that the model encodes useful information at the phoneme level.

In the second experiment, for some words such as 'person' or 'men', word recognition was highest after the first phoneme, and decreased upon more phonemes being presented to the model. This can partly be explained by the results seen in the first experiment, where it was shown that words that sound similar can have similar embeddings. Likely, when more phonemes of the tested words became available to the model, the words became more phonemically similar to word competitors, which activated the representations of competing words instead of the representations of the tested words.

Furthermore, in the second experiment, words were presented in phoneme sequences of increasing length both from word onset and from word offset. This experiment showed that the VGS model recognised words better when they were presented from word onset than from word offset. However, the VGS model was also able to recognise words

when presented from word offset without having been exposed to the initial part of a word, which is in line with what has been shown in gating experiments to happen with human listeners [36]. This is different from the findings by Havard and colleagues, as they found that word onset was needed for word recognition in their VGS model [12]. This difference in findings may have been caused by the fact that this research employs a bi-directional GRU in the speech encoder, while Havard and colleagues employed a uni-directional GRU, which only learns speech time-dependencies in a single direction [12].

The third experiment showed that word recognition could be improved by 'priming' the GRU of the model with the caption preceding a word token. As mentioned in Section 4.4, the increase in word recognition performance was strongest for words which were in the bottom 50th percentile with regard to 'unprimed' word recognition scores, with a relative performance increase of 27% reached by 'priming' the GRU of the VGS model with the preceding caption. Human word recognition speed can also be improved by priming a human listener with a semantically related word (Section 2.1.3). Although providing the VGS model with a preceding caption is not precisely semantic priming, it does show how providing contextual information which could be semantically related to a word can substantially improve word recognition for some words.

5.2. LIMITATIONS

In this thesis, a new method for evaluating word recognition in VGS models is adopted from [12], by investigating word recognition through retrieving image embeddings with a high cosine similarity to segmented word embeddings. However, this method does entail some limitations, as word recognition is evaluated through image retrieval scores, while using textual image captions as ground-truth labels for word recognition. As a result, word recognition scores often do not reflect how well the model actually learned the meaning of certain words. For example, the model recognises the word 'man' really well, while the word 'men' was often not recognised, although the images that were retrieved often contained a man or multiple men. Similarly, for the words 'face', 'hair', 'mouth' and 'head', the model often retrieved an image with a human head in it, while the ground-truth caption often contained words like 'person', 'man', or 'woman'.

As the above examples illustrate, the captions which accompany the image do not always serve as a perfect ground-truth value for word recognition, as they do not fully encompass the image content. Ideally, if word recognition is analysed, images should be annotated with a set number of classes which fully encompass the content of an image. There are two ways to potentially improve on this. Firstly, this could be resolved by using a multi label classification network to create labels for the words present in the images in the Flickr8k dataset, and using those labels to evaluate word recognition. Secondly, a dataset of images with annotated classes could be used, however those datasets would most likely not contain naturally spoken captions, and would instead require the use of synthetically spoken captions and words.

Furthermore, due to the fact that VGS models are trained on pairs of captions and images, factors within the images might also impact the word recognition performance of a VGS model. The effect of image factors on word recognition were left out of the scope of this research for a number of reasons. Firstly, the ResNet-152 image encoder used in

this research was originally trained for one-class image classification, and as a result did not allow for the construction of image factors such as the number or size of objects in the images. Secondly, the Flickr8k dataset does not seperately provide extra image information such as object segmentation or information about the objects present within an image.

6

Future Work

As discussed in Section 5.2, a VGS model is trained on pairs of captions and images. As a result, factors within the images might also impact the word recognition performance of a VGS model. Image factor effects could be investigated by using the MSCOCO dataset. Although this dataset would require the use of synthetic speech captions, it does contain object segmentation information (information about the location of an object in an image) and information on the number of objects present in an image as well as ground-truth object labels. As a result, employing such a dataset could give insight into the role of images during the learning process of VGS models.

Furthermore, using a dataset which contains image labels could also allow a more extensive analysis of misclassifications during word recognition. The dataset used in this research does not contain ground-truth labels for its images, which is why it is difficult to determine in which way words are misclassified if they if they are not recognised. This research has found that words are embedded both on semantic meaning and on the units of sound present in the words, which potentially causes some misclassifications for words that are semantically similar or sound similar. Having ground-truth labels could reveal more information on how a VGS model misclassifies words, and as a result how a model learns to recognise words.

Also, there might be more linguistic and acoustic factors affecting word recognition in VGS models besides the ones addressed in this research. It would for example be interesting to research the effect of other factors, such as pitch and volume of a speech signal. Likely, variation in the speech signals caused by these factors could also have an influence on word recognition.

Building on the findings of the third experiment, it would be interesting to further explore 'priming' effects within VGS models. Specifically, it would be interesting to see whether cross-modal priming experiments could be performed, where a word is primed with a visual stimulus. Such an experiment could give insight into how the alignment between speech and images is made by the VGS model.

Furthermore, it would be interesting to see if certain factors, which were found to affect word recognition in this research, could be used to improve word recognition for VGS models trained on captions and images. For example, the effect of the frequency of a word (in the training set) on word recognition demonstrated how reliant a VGS model is on the word distribution of its training data. These word frequency effects could potentially be eliminated by pre-training the network on a dataset of synthetic speech captions and images which is tailored to have an even distribution of word occurrence. This would provide the VGS model with better exposure to less frequent words, before being fine-tuned on another dataset with naturally spoken captions such as Flickr8k. Similarly, Ilharco and colleagues pre-trained their VGS model on a large dataset of synthetic images and captions, before fine-tuning on the Flickr8k dataset, which resulted in them achieving state-of-the-art performance on the caption-to-image retrieval task [10].

Also, it would be interesting to see whether certain architectural changes, if made to the VGS model used in this research, could allow it to learn better from the data it is trained on. For example, LSTM units could potentially be implemented to replace the GRU units currently used, as they have shown to consistently outperform GRUs on many classification tasks [82]. GRUs however have shown to outperform LSTMs on certain

smaller datasets [83], such as Flickr8k. Potentially, when the VGS architecture used in this research is applied to a larger dataset of images and captions, LSTM units could help increase performance on both the caption-to-image retrieval and word recognition tasks.

Lastly, in the more distant future, the 'holy grail' of VGS models would be if a mechanism could be developed which allowed such a model to learn in a truly unsupervised manner. Currently, the models are trained on a curated dataset of image-caption pairs. If VGS models could be developed to work in a truly unsupervised manner and would 'know' how to discriminate semantically related image-caption pairs to train on from non related pairs, they could potentially be trained to create a visual-semantic alignment on large amounts of unordered data, such as movies. This could eliminate the need for curating training data in order to develop speech technologies using a VGS model.

7

Conclusion

THIS research investigated whether and how words are recognised by a visually grounded speech model using real human speech. Based on the results of the experiments, the research questions can be answered as follows.

- Research Question 1: *Can a Visually Grounded Speech model perform word recognition with naturally spoken speech?*

The first experiment revealed that a VGS model is able to perform word recognition using naturally spoken speech. This suggests that, during training time, the VGS model implicitly learned individual words from continuous speech, while not being explicitly supervised to do so. The VGS model was able to recognise words in 4.4 out of 10 cases. These results indicate that the relatively 'optimistic' results shown in [12] do not accurately reflect the difficulty of the word recognition task, and that naturally spoken words pose a bigger challenge for word recognition than synthetically spoken captions.

- Research Question 2: *How are naturally spoken words recognised by a Visually Grounded Speech model?*

The second research question is answered through a number of sub questions defined in the problem statement.

- Sub Question 1: *What is the time-course of word recognition?*

In the second experiment word tokens segmented at increasing length from both word onset and word offset were presented to the model. This experiment showed that words are recognised better when presented from word onset than from word offset. When words are presented from word onset, word recognition generally improves the most while the first 50% of a word's phonemes are being presented to the model. Furthermore, it showed that the steepest increase in word recognition happens after the VGS model is presented with 30-40% of a word's phonemes from word onset. When words are presented from word offset, word recognition improves considerably when the model is presented with 60-100% of a word's phonemes. When words are presented from word offset, recognition increases the steepest after 80-90% of a word's phonemes are presented to the model.

- Sub Question 2: *What is the amount of information needed for word recognition?*

The second experiment showed that the VGS model recognises most words after having been exposed to only 30-40% of a word's phonemes when presented from word onset. To achieve a similar performance from word offset, the model would need 60-70% of a word's phonemes. Some words are not recognised at all regardless of the amount of speech information provided to the model. As explained in Section 5.2, this is likely due to the image labels in this research not always accurately representing the image content.

- Sub Question 3: *Is the model able to encode units of sound, such as words and phonemes?*

The first experiment revealed that the VGS model implicitly learns words from the speech captions it was trained on, and as a result is able to recognise words. This experiment also showed that the model encoded both acoustic information and semantic meaning in its word embeddings. The second experiment revealed that the VGS model also implicitly segments phonemes, and that the model can already recognise words by being exposed to a partial sequence of a word's phonemes.

- Sub Question 4: *How does contextual information affect word recognition?*

The third experiment revealed that word recognition can be enhanced by 'priming' the model with a word token's preceding caption. Word types that were generally not recognised well, and belonged to the lowest 50th percentile in P@10 scores, turned out to benefit the most from this contextual information, and saw a relative increase in word recognition performance of 27%.

- Sub Question 5: *What linguistic and acoustic factors affect word recognition?*

The statistical analyses performed using the LMER in the first and second experiments revealed a number of linguistic and acoustic factors which affect word recognition. The most notable factors affecting word recognition were: speaking rate, signal duration, training set word frequency and the size of the word-initial cohort.

This research demonstrated that a VGS model which is trained on naturally spoken speech and images can be used to perform word recognition. The VGS model learned individual words from continuous speech, while not being explicitly supervised to do so. Furthermore, it learned to capture both acoustic information and semantic meaning in its word embeddings. This research presents a methodology for investigating how word recognition is taking place in VGS models by taking inspiration from human speech recognition research. Through this methodology, it was found that words can already be recognised by being exposed to only a partial sequence of a word's phonemes. Furthermore, it was found that word recognition can be enhanced by 'priming' the VGS model with contextual information. Lastly, this research showed that some factors which affect word recognition in humans also seem to affect word recognition in VGS models, in some cases in a similar way. The results of - and methodology used in - this research can contribute to a better understanding of how words are learned in VGS models, and consequently to the development of better speech technologies using such models for unwritten, low and high-resource languages.

8

References

REFERENCES

[1] O. Räsänen and H. Rasilo, *A joint model of word segmentation and meaning acquisition through cross-situational learning.* Psychological review **122**, 792 (2015).

[2] M. Tomasello, *The usage-based theory of language acquisition,* in *The Cambridge handbook of child language* (Cambridge Univ. Press, 2009) pp. 69–87.

[3] D. Harwath and J. Glass, *Deep multimodal semantic embeddings for speech and images,* in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (IEEE, 2015) pp. 237–244.

[4] D. Harwath, A. Torralba, and J. Glass, *Unsupervised learning of spoken language with visual context,* in *Advances in Neural Information Processing Systems* (2016) pp. 1858–1866.

[5] G. Chrupala, L. Gelderloos, and A. Alishahi, *Representations of language in a model of visually grounded speech signal,* in *Proceedings of the 55th of the Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2017) p. 613–622.

[6] D. Merkx, S. L. Frank, and M. Ernestus, *Language learning using speech to image retrieval,* in *Proceedings of Interspeech 2019. Crossroads of Speech and Language* (2019).

[7] H. Kamper, G. Shakhnarovich, and K. Livescu, *Semantic keyword spotting by learning from images and speech,* arXiv preprint arXiv:1710.01949 (2017).

[8] O. Scharenborg, L. Besacier, A. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, M. Müller, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang, and E. Dupoux, *Speech technology for unwritten languages,* IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 964 (2020).

[9] H. Kamper and M. Roth, *Visually grounded cross-lingual keyword spotting in speech,* The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (2018).

[10] G. Ilharco, Y. Zhang, and J. Baldridge, *Large-scale representation learning from visually grounded untranscribed speech,* arXiv preprint arXiv:1909.08782 (2019).

[11] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh, *Symbol emergence in robotics: a survey,* Advanced Robotics **30**, 706 (2016).

[12] W. N. Havard, J.-P. Chevrot, and L. Besacier, *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) pp. 339–348.

[13] D. Harwath, W.-N. Hsu, and J. Glass, *Learning hierarchical discrete linguistic units from visually-grounded speech,* arXiv preprint arXiv:1911.09602 (2019).

[14] G. Chrupała, B. Higy, and A. Alishahi, *Analyzing analytical methods: The case of phonology in neural models of spoken language,* arXiv preprint arXiv:2004.07070 (2020).

[15] W. Marslen-Wilson, *Activation, competition, and frequency in lexical access.* (1990).

[16] E. Weingarten, Q. Chen, M. McAdams, J. Yi, J. Hepler, and D. Albarracín, *From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words.* Psychological Bulletin **142**, 472 (2016).

[17] O. Scharenborg, *Reaching over the gap: A review of efforts to link human and automatic speech recognition research,* Speech Communication **49**, 336 (2007).

[18] A. Weber and O. Scharenborg, *Models of spoken-word recognition,* Wiley Interdisciplinary Reviews: Cognitive Science **3**, 387 (2012).

[19] O. Scharenborg and L. Boves, *Computational modelling of spoken-word recognition processes: Design choices and evaluation,* Pragmatics & Cognition **18**, 136 (2010).

[20] W. D. Marslen-Wilson and A. Welsh, *Processing interactions and lexical access during word recognition in continuous speech,* Cognitive psychology **10**, 29 (1978).

[21] R. A. Cole, *Listening for mispronunciations: A measure of what we hear during speech,* Perception & Psychophysics **13**, 153 (1973).

[22] M. Taft and G. Hambly, *Exploring the cohort model of spoken word recognition,* Cognition **22**, 259 (1986).

[23] J. L. McClelland and J. L. Elman, *The trace model of speech perception,* Cognitive psychology **18**, 1 (1986).

[24] D. Dahan, J. S. Magnuson, and M. K. Tanenhaus, *Time course of frequency effects in spoken-word recognition: Evidence from eye movements,* Cognitive psychology **42**, 317 (2001).

[25] D. Norris, *Shortlist: A connectionist model of continuous speech recognition,* Cognition **52**, 189 (1994).

[26] P. A. Luce and M. S. Cluff, *Delayed commitment in spoken word recognition: Evidence from cross-modal priming,* Perception & Psychophysics **60**, 484 (1998).

[27] C. P. Whaley, *Word—nonword classification time,* Journal of Verbal learning and Verbal behavior **17**, 143 (1978).

[28] M. S. Sommers, L. C. Nygaard, and D. B. Pisoni, *Stimulus variability and spoken word recognition. i. effects of variability in speaking rate and overall amplitude,* The Journal of the Acoustical Society of America **96**, 1314 (1994).

[29] R. A. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey, *The contribution of consonants versus vowels to word recognition in fluent speech,* in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 2 (IEEE, 1996) pp. 853–856.

[30] D. Norris, J. M. McQueen, and A. Cutler, *Competition and segmentation in spoken-word recognition.* Journal of Experimental Psychology: Learning, Memory, and Cognition **21**, 1209 (1995).

[31] L. K. Tyler, *The structure of the initial cohort: Evidence from gating,* Perception & Psychophysics **36**, 417 (1984).

[32] L. M. Slowiaczek, H. C. Nusbaum, and D. B. Pisono, *Phonological priming in auditory word recognition.* Journal of Experimental Psychology: Learning, Memory, and Cognition **13**, 64 (1987).

[33] B. Munson and N. P. Solomon, *The effect of phonological neighborhood density on vowel articulation,* Journal of speech, language, and hearing research (2004).

[34] F. Grosjean, *Spoken word recognition processes and the gating paradigm,* Perception & psychophysics **28**, 267 (1980).

[35] W. Marslen-Wilson and L. K. Tyler, *The temporal structure of spoken language understanding,* Cognition **8**, 1 (1980).

[36] F. Grosjean, *The recognition of words after their acoustic offset: Evidence and implications,* Perception & Psychophysics **38**, 299 (1985).

[37] P. R. Chiappe, M. C. Smith, and D. Besner, *Semantic priming in visual word recognition: Activation blocking and domains of processing,* Psychonomic Bulletin & Review **3**, 249 (1996).

[38] J. H. Neely, *Semantic priming effects in visual word recognition: A selective review of current findings and theories,* Basic processes in reading: Visual word recognition **11**, 264 (1991).

[39] B. H. Juang and L. R. Rabiner, *Hidden markov models for speech recognition,* Technometrics **33**, 251 (1991).

[40] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, *Convolutional neural networks for speech recognition,* IEEE/ACM Transactions on audio, speech, and language processing **22**, 1533 (2014).

[41] A. Graves, A.-r. Mohamed, and G. Hinton, *Speech recognition with deep recurrent neural networks,* in *2013 IEEE international conference on acoustics, speech and signal processing* (IEEE, 2013) pp. 6645–6649.

[42] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. K. Chin, *et al.*, *Acoustic modeling for google home.* in *Interspeech* (2017) pp. 399–403.

[43] P. Dayan, M. Sahani, and G. Deback, *Unsupervised learning,* The MIT encyclopedia of the cognitive sciences , 857 (1999).

[44] J. Pan, C. Liu, Z. Wang, Y. Hu, and H. Jiang, *Investigation of deep neural networks (dnn) for large vocabulary continuous speech recognition: Why dnn surpasses gmms in acoustic modeling,* in *2012 8th International Symposium on Chinese Spoken Language Processing* (IEEE, 2012) pp. 301–305.

[45] A. Graves and N. Jaitly, *Towards end-to-end speech recognition with recurrent neural networks,* in *International conference on machine learning* (2014) pp. 1764–1772.

[46] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition,* Proceedings of the IEEE **86**, 2278 (1998).

[47] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagating errors,* nature **323**, 533 (1986).

[48] S. W. Smith *et al., The scientist and engineer's guide to digital signal processing,* (1997).

[49] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, *Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition,* in *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)* (IEEE, 2012) pp. 4277–4280.

[50] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, *Speech recognition using deep neural networks: A systematic review,* IEEE Access **7**, 19143 (2019).

[51] S. Hochreiter and J. Schmidhuber, *Long short-term memory,* Neural computation **9**, 1735 (1997).

[52] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, *Learning phrase representations using rnn encoder-decoder for statistical machine translation,* arXiv preprint arXiv:1406.1078 (2014).

[53] I. Sutskever, O. Vinyals, and Q. V. Le, *Sequence to sequence learning with neural networks,* in *Advances in neural information processing systems* (2014) pp. 3104–3112.

[54] D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate,* arXiv preprint arXiv:1409.0473 (2014).

[55] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, *Attention-based models for speech recognition,* in *Advances in neural information processing systems* (2015) pp. 577–585.

[56] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, *End-to-end attention-based large vocabulary speech recognition,* in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE, 2016) pp. 4945–4949.

[57] U. Shrawankar and V. M. Thakare, *Techniques for feature extraction in speech recognition system: A comparative study,* arXiv preprint arXiv:1305.1145 (2013).

[58] M. P. Kesarkar, *Feature extraction for speech recognition,* Electronic Systems, EE. Dept., IIT Bombay (2003).

[59] N. Dave, *Feature extraction methods lpc, plp and mfcc in speech recognition,* International journal for advance research in engineering and technology **1**, 1 (2013).

[60] P. Mermelstein, *Distance measures for speech recognition, psychological and instrumental,* Pattern recognition and artificial intelligence **116**, 374 (1976).

[61] L. Muda, M. Begam, and I. Elamvazuthi, *Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques,* arXiv preprint arXiv:1003.4083 (2010).

[62] A. Hagen, D. A. Connors, and B. L. Pellom, *The analysis and design of architecture systems for speech recognition on modern handheld-computing devices,* in *First IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and Systems Synthesis (IEEE Cat. No. 03TH8721)* (IEEE, 2003) pp. 65–70.

[63] F. Zheng, G. Zhang, and Z. Song, *Comparison of different implementations of mfcc,* Journal of Computer science and Technology **16**, 582 (2001).

[64] S. Wijoyo and S. Wijoyo, *Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile robot,* in *Proceedings of 2011 International Conference on Information and Electronics Engineering (ICIEE 2011)* (2011) pp. 28–29.

[65] A. Karpathy and L. Fei-Fei, *Deep visual-semantic alignments for generating image descriptions,* in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015) pp. 3128–3137.

[66] H. Kamper, G. Shakhnarovich, and K. Livescu, *Semantic speech retrieval with a visually grounded model of untranscribed speech,* IEEE/ACM Trans. Audio, Speech and Lang. Proc. **27**, 89–98 (2019).

[67] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, *Jointly discovering visual objects and spoken words from raw sensory input,* in *Proceedings of the European conference on computer vision (ECCV)* (2018) pp. 649–665.

[68] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *ImageNet: A Large-Scale Hierarchical Image Database,* in *CVPR09* (2009).

[69] R. Girshick, J. Donahue, T. Darrell, and J. Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014) pp. 580–587.

[70] A. Alishahi, M. Barking, and G. 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).

[71] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition,* arXiv preprint arXiv:1409.1556 (2014).

[72] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 770–778.

[73] J. G. Zilly, R. K. Srivastava, J. Koutník, and J. Schmidhuber, *Recurrent highway networks,* arXiv preprint arXiv:1607.03474 (2016).

[74] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization,* arXiv preprint arXiv:1412.6980 (2014).

[75] L. N. Smith, *Cyclical learning rates for training neural networks,* in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, 2017) pp. 464–472.

[76] M. Hodosh, P. Young, and J. Hockenmaier, *Framing image description as a ranking task: Data, models and evaluation metrics,* Journal of Artificial Intelligence Research **47**, 853 (2013).

[77] D. Bates, M. Mächler, B. Bolker, and S. Walker, *Fitting linear mixed-effects models using lme4,* Journal of Statistical Software **67**, 1 (2015).

[78] D. Bates, M. Mächler, B. Bolker, and S. Walker, *Fitting linear mixed-effects models using lme4,* arXiv preprint arXiv:1406.5823 (2014).

[79] M. S. Vitevitch and P. A. Luce, *Phonological neighborhood effects in spoken word perception and production,* Annual Review of Linguistics **2**, 75 (2016).

[80] S. Scholten, D. Merkx, and O. Scharenborg, *Learning to recognise words using visually grounded speech,* arXiv preprint arXiv:2006.00512 (2020).

[81] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft coco: Common objects in context,* in *European conference on computer vision* (Springer, 2014) pp. 740–755.

[82] D. Britz, A. Goldie, M.-T. Luong, and Q. Le, *Massive exploration of neural machine translation architectures,* arXiv preprint arXiv:1703.03906 (2017).

[83] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, *Empirical evaluation of gated recurrent neural networks on sequence modeling,* arXiv preprint arXiv:1412.3555 (2014).

A

Appendix A

Learning to Recognise Words using Visually Grounded Speech

Sebastiaan Scholten[1], *Danny Merkx*[2], *Odette Scharenborg*[1]

[1]Multimedia Computing Group, Delft University of Technology, Delft, the Netherlands
[2]Centre for Language Studies, Radboud University, Nijmegen, the Netherlands
J.S.M.Scholten@student.tudelft.nl, D.Merkx@let.ru.nl, O.E.Scharenborg@tudelft.nl

Abstract

We investigated word recognition in a Visually Grounded Speech model. The model has been trained on pairs of images and spoken captions to create visually grounded embeddings which can be used for speech to image retrieval and vice versa. We investigate whether such a model can be used to recognise words by embedding isolated words and using them to retrieve images of their visual referents. We investigate the time-course of word recognition using a gating paradigm and perform a statistical analysis to see whether well known word competition effects in human speech processing influence word recognition. Our experiments show that the model is able to recognise words, and the gating paradigm reveals that words can be recognised from partial input as well and that recognition is negatively influenced by word competition from the word initial cohort.

Index Terms: Visually Grounded Speech, Recurrent Neural Network, Flickr8k, Analysis.

1. Introduction

Babies initially have little semantic understanding of what is being said around them. It is theorized that the fact that they repeatedly hear certain words while they observe certain objects around them enables them to learn a mapping between speech and objects [1]. Repetitive hearing of these utterances in the context of some functional consistency, such as picking up an object, will display the meaning of a smaller constituent of such an utterance, e.g., a word, and potentially about the class of objects it belongs to [2].

Some core principles of Visually Grounded Speech (VGS) models are inspired by such learning processes. While most speech recognition research focuses on speech signals only, Visually Grounded Speech models include visual information rather than textual transcriptions to guide the training of the acoustic models [3, 4, 5, 6, 7, 8, 9] . Following the approach of multimodal neural models which produce visual-semantic alignments for images and text [10], a VGS model employs two parallel Deep Neural Networks (DNNs) which are trained to map a speech signal and a corresponding image into a common embedding space.

Recent research on VGS models has seen an improvement in architectures and training schemes [5, 11, 6] and different applications of the VGS model have been proposed such as semantic keyword spotting [7, 12] and speech-based image retrieval [3, 5, 6, 8], where a trained VGS model is fed full sentence speech captions with which the model retrieves the corresponding image. Recent research has shown that VGS models implicitly learn to recognise meaningful sentence constituents such as phonemes and words and the presence of these constituents can be decoded from the speech embeddings [5, 6, 13, 14, 15]. Havard and colleagues presented isolated words to a VGS model and investigated whether the model was able to retrieve images of the words' correct visual referents [13]. This showed that the model does not just encode these constituents into the speech embeddings, but the model actually 'recognises' individual words and learned to map them onto their correct visual referents.

Building on the synthetic speech experiments by Havard and colleagues, we investigate how natural speech is recognised by a VGS model using real human speech. In this paper, we will 1) investigate isolated word recognition using real speech, 2) investigate how words are recognised by a VGS model over time, 3) and look more in depth into the linguistic and acoustic properties that aid or hinder word recognition. As in [13], we use the retrieval of images containing a word's correct visual referent as a measure of the model's word recognition performance. Word recognition is expected to be more challenging with real speech as opposed to synthetic speech, due to real speech having more variation in quality, noise and speaking rate synthetic speech. This can also be seen in [5], where the model trained on real speech performs significantly worse with caption-to-image retrieval than a model trained on synthetic speech.

We carry out two experiments, inspired by those of [13]. In our first experiment, the VGS model is fed individual words, which will allow us to investigate whether the model is actually learning to recognise individual words, which would be shown by the model being able to retrieve a relevant image on the basis of a single word rather than the full caption. In the second experiment, we use a gating paradigm, borrowed from human speech processing research. In the gating experiment, a word is presented to the VGS model in speech segments of increasing duration, i.e., with increasing number of phones, and 'asked' to retrieve an image of the correct visual referent on the basis of the available phone string. This allows us to investigate 1) the time-course of word recognition, 2) the amount of information needed for word recognition, and 3) whether the model is able to encode phones in the combined embedding space.

To answer our third question, we carry out a statistical analysis in which word recognition performance is predicted using several linguistic and acoustic features. These linguistic and acoustic features are factors known to influence human speech processing. In human speech processing (see for an overview of models of human speech processing Weber & Scharenborg [16]), the incoming speech signal is mapped against phone representations in the listener's brain, and the sounds that best resemble the incoming speech signal are 'activated'. These activated phone representations, activate every possible word in which they appear, irrespective of the position of the phone in the word. As more speech information becomes available, words that no longer match the input will drop out of the list of activated words. The word that best matches the speech input is recognised. Words that are activated are called competitors or competitor words. The number of competitor words plays a role in human speech processing: the more competitors there

are, the longer it takes for a word to be recognised [17]. We want to see whether our VGS model activates competitor words in a similar manner, which would be shown by a significant effect of the number of competitor words on word recognition performance. We focus on the number of words that share the start of the word, the so-called word-initial cohort, as we are testing isolated words in our experiment [18], and the neighbourhood density, i.e., the number of words that differs exactly one phoneme from the target word.

The rest of this paper is organised as follows. Firstly, we discuss the model architecture and the methodology behind the experiments. Secondly, the results for the different experiments will be discussed. Lastly, this work will be concluded with a discussion with a summary of the contributions, as well as recommendations for future research.

2. Methodology

2.1. Visually Grounded Speech Model

For this paper, we use the Visually Grounded Speech Model implementation presented in [6], with the addition of an extra Gated Recurrent Unit (GRU) layer, which can improve the the model's ability to capture long-range dependencies. The model consists of two DNNs: a pretrained image encoder and a Recurrent Neural Network (RNN)-based speech caption encoder. The encoders embed the speech and images, and the model is trained to minimise the cosine distance between image-caption pairs in the shared embedding space. A visual representation of the model is given in Figure 1.

The pre-trained image encoder is ResNet-152, which was trained on ImageNet [19]. The final object classification layer is removed from this network, and we place a single linear layer on top of ResNet and L2 normalise the result to map the latent image features onto our multimodal embedding space.

Our audio features consist of Mel Frequency Cepstral Coefficients (MFCCs). A 39-dimensional feature vector was used, comprising of 12 MFCCs including their log energy feature and first and second derivatives. A 1-dimensional convolutional layer was applied to the 39-dimensional feature vector, then these channels were fed to an RNN with a 4-layer bi-directional GRU. Then, the 1024 bi-directional units were concatenated to create a 2048 feature vector, which feeds into a self-attention layer. The resulting feature representations are L2 normalised to arrive at the final caption embedding.

The caption encoder was trained in order for the image and speech pairs to have a cosine similarity larger by a margin α than the cosine similarity for mismatched pairs. We used a hinge loss function to minimise cosine distance for ground-truth pairs. The model was trained for 32 epochs with a batch size of 32. For a more detailed description of the model and the loss function please refer to [6].

We train the model on Flickr8k [20], a database with 8k images and 5 written captions per image for a total of 40k captions. Harwath and colleagues collected spoken versions of these captions from a total of 183 different speakers, with a vocabulary of 8918 unique words [3]. For our training, validation and test set we make use of the data split provided by [10]. We use spoken caption-to-image retrieval to evaluate how well our model performs on the training task and compare the model with previous work. Caption-to-image retrieval is measured in Recall@N, the percentage of captions for which the correct image was in the top N retrieved images. Images are retrieved based on their embedding distance to the caption embedding.

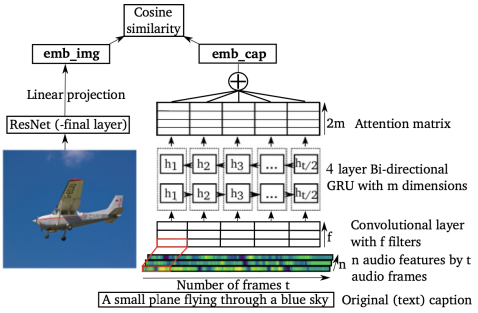


Figure 1: *A visual representation of the image encoder parallel to the caption encoder. Based on [6].*

2.2. Experiments

We will be performing two experiments. In the first experiment, we present our model with isolated words to investigate how well the model learned to map these words onto their visual referents. In the second experiment, we segment the words into phonemes and present our model with phoneme sequences of increasing length to investigate the time-course of word recognition in the model. We present our model with multiple instances of each word, spoken by different speakers to gain a more realistic impression of how a word performs across different speakers and contexts. Also, this allows us to test which acoustic factors in the speech signal are influencing the model's word recognition performance.

2.2.1. Experimental data

A visually grounded model relies on there being a consistency between the image and speech signal in order to create a common embedding space. Therefore, we chose 49 words with clear visual referents, such as 'bike' and 'man', as opposed to articles and adverbs. We extracted 50 occurrences of each word from the speech captions in the test set, to have an equal sample size for each word to allow a fair comparison between their word recognition performance.

The words were extracted from the speech signal using a forced alignment of the phonetic transcriptions with the speech captions in Flick8k. For the second experiment, these words were segmented into sequences of phonemes where each sequence was one phoneme longer than the previous. For example, for the word 'bike', the speech signal was segmented into 'B', 'B-AY', and 'B-AY-K'.

2.2.2. Evaluating word recognition performance

Following [13], we use the retrieval of images containing a word's correct visual referent as a measure of the model's word recognition performance. In order to quantify this we use the Precision@10 score which is calculated as follows. We use the trained VGS model to create embeddings for all of the word instances. From the Flickr8k test set, we take all images which had one of our 49 words in its captions and use the VGS model to create image embeddings. For each embedded word instance we then retrieve the ten most similar image embeddings as defined by cosine similarity between the embeddings. The Precision@10 (P@10) is then calculated for each word instance as

the percentage of its top ten images which contain the correct visual referent of the word.

2.2.3. Evaluating linguistic and acoustic factors

To answer our third research question, we examine linguistic and acoustic factors which might influence the model's word recognition performance using a Linear Mixed Effects Regression (LMER). For the LMER analysis we used the *lme4* package in R [21]. All fixed effects are z-score normalised. The dependent variable is the P@10 score.

For the word recognition experiment, our LMER model takes into consideration the signal duration (i.e., number of speech frames), the speaking rate calculated as the number of phonemes in the word divided by its signal duration, the frequency of occurrence of the word in the training set and the number of phonemes, vowels, and consonants in the word. We also included the two-way interaction of the frequency of occurrence of the word in the training set with the number of phonemes, vowels, and consonants. We considered these interaction effects because words with a certain number of phonemes, vowels, and consonants might appear more often in a dataset. Furthermore, we included by-speaker and by-word random intercepts and by-speaker random slopes for the signal length, to take into consideration speaker differences on the duration of the signal.

For the second experiment, the LMER model takes into account the earlier mentioned frequency of occurrence of the word in the training set and the total number of phonemes in the word. We also include the size of the word-initial cohort and neighbourhood density. The word-initial cohort is calculated by determining for each phoneme sequence the number of words which start with the same phoneme sequence in the Flickr8k training set, which considers a total of 6182 unique words. This indicates the number of words that is considered simultaneously for recognition by the model given the phoneme sequence seen so far. The neighbourhood density is calculated as the number of words from the words in the Flickr8k training set that can be formed from the phoneme sequence by a one-phoneme substitution [22]. This factor indicates the similarity among spoken forms of words, and is therefore a second measure of the number of words that are simultaneously considered for recognition. The model also includes a by-speaker and a by-word random intercept.

3. Results

The scores in Table 1 show the result for the speech caption-to-image retrieval task. This indicates how well the model learned to embed the speech and images in the common embedding space. R@N is the percentage of items for which the correct image was in the top N retrievals. Median R is the median rank of the correctly retrieved image. The addition of an extra GRU layer has led to a substantial performance increase, allowing dependencies in longer speech captions to be captured better.

Table 1: *Speech caption-to-image retrieval scores including 95% confidence intervals for our model. For comparison, the models of Merkx et al. [6], Chrupała et al. [5] and Harwath et al. [3] which were also trained on Flickr8k speech captions are provided.*

Model	R@1	R@5	R@10	Med. R
4-GRU	10.71±1.9	29.2±2.8	40.2±3.0	18
[6]	8.0±1.7	24.5±2.7	35.5±3.0	24
[5]	5.5±1.4	16.3±2.3	25.3±2.7	48
[3]			17.9±2.4	

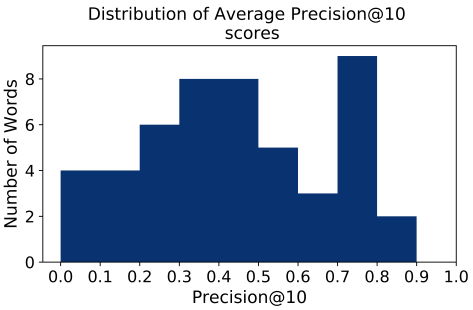


Figure 2: *Distribution of Average P@10 scores for the 49 tested words, assigned to bin intervals of size 0.1.*

3.1. Word recognition

In this experiment, we present isolated words to the model. The histogram in Figure 2 shows the distribution of the P@10 scores over the 49 words. The average P@10 is 0.44, which indicates that on average 4.4 out of the ten retrieved images contain the correct visual referent. However, Figure 2 also shows that four words have a P@10 near zero, meaning that no correct images were retrieved and the word was not recognised. Furthermore, Havard and colleagues [13] reported a median P@10 of 0.8, while we on the other hand have a median P@10 of 0.4. While our model does learn to recognise most words to some degree, this indicates a large difference in recognition performance going from the synthetic speech dataset in [13] to the real speech of Flickr8k.

Table 2 shows the results from the statistical test. Firstly, signal duration was found to have a significant negative effect on the P@10 scores. This shows that the model has more difficulty encoding longer words. Secondly, speaking rate also had a significant negative effect, showing that words that are spoken more rapidly were encoded less well than words pronounced more slowly. Lastly, the frequency of occurrence of the word in the training set was shown to have a significant positive effect on word recognition performance. This shows that words which occur more often in training samples are encoded considerably better for word recognition. No interaction effects were found.

For our random effects, we see that the standard deviation of the scores between words is far larger than between speakers.

Table 2: *Significant fixed effects with Standard Errors for the word recognition LMER.*

Fixed effects	Estimate	P-value
Intercept	0.432±0.033	<0.001
Signal duration	-0.050±0.014	<0.001
Speaking rate	-0.068±0.013	<0.001
Training set frequency	0.152±0.063	0.020

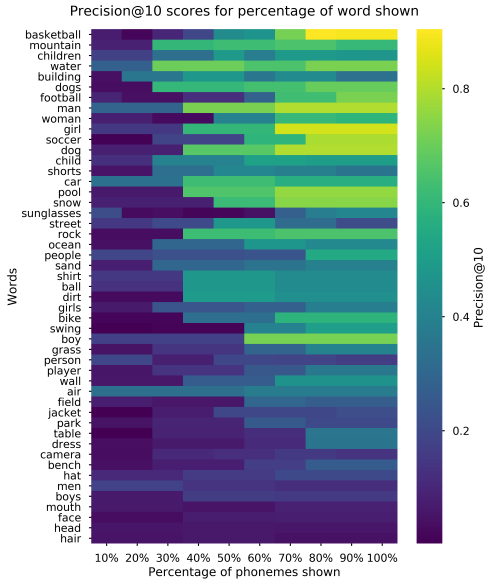


Figure 3: *Heatmap showing the P@10 scores of a given word (shown on the y-axis) as a function of the phoneme sequence length. The x-axis indicates the percentage of phonemes of the word that were available to the model.*

This shows that the effect of using different speakers causes less variation in results in comparison to using different words.

3.2. Word activation

In order to investigate the time-course of word recognition and how much information is needed for word recognition, phoneme sequences of increasing length were given to the model. Figure 3 shows the results in terms of the P@10 of a given word (shown on the y-axis) as a function of the phoneme sequence length in terms of percentage of phonemes of the word. Note that the x-axis has ten values, if a word has for instance only two phonemes, the P@10 for the first and second phoneme span 10-50% and 60-100% respectively. A more yellow colour corresponds to a higher P@10.

As can be seen in Figure 3, generally, the more phonemes of a word the model is exposed to, the better it can retrieve the image corresponding to the spoken word. Some words representations, see the bottom of Figure 3 (bars are entirely blue), are not recognised at all irrespective of the percentage of phonemes shown to the model.

The results of the LMER model are summarised in Table 3. Unsurprisingly, the number of phonemes in a phoneme sequence has a significant positive effect on the P@10 scores indicating that words are recognised better when the model is presented with longer phoneme sequences. The frequency of occurrence of the word in the training set again has a significant positive effect on the performance, showing that having more training examples allows phoneme sequences to be mapped more easily to the correct visual referent. The word-initial cohort has a significant negative effect on the P@10 scores, indicating that, similar to human listeners, word recognition is more difficult when there are more words that have the same phoneme sequence at the start of the word. The effect of the neighbourhood density was not found to be significant.

Table 3: *Significant fixed effects with Standard Errors for the word activation LMER.*

Fixed effects	Estimate	P-value
Intercept	0.295±0.020	<0.001
# of phonemes	0.134±0.003	<0.001
Training set frequency	0.087±0.018	<0.001
Word-initial cohort	-0.037±0.003	<0.001

4. Discussion and Conclusions

In this paper, we investigated how natural speech is recognised by a Visually Grounded Speech model using real human speech. In order to do this, in the first experiment, we investigated how isolated words are recognized in a VGS model. Although our model is trained on full speech captions, the word recognition experiment showed that the model learned to recognise individual words and was able to map them onto their correct visual referent in most cases.

Also, we investigated the time course of the word recognition. The second experiment showed that it is possible to recognise a word from only a partial phoneme sequence and that word recognition performance (as measured in image retrieval scores) generally improved as more phonemes were seen, with the best retrieval scores when the model was shown all phonemes of the word. The largest leap in word recognition performance was observed after the model was provided with a phoneme sequence consisting of 30%-40% of the target word's phonemes. For some words such as 'person' or 'men', word recognition was highest right after the first phoneme and decreased upon seeing more of the speech signal, although in these cases the word generally was not recognised well. Similar to human listeners [16], the model did not need to have available all phonemes of the word in order to recognize it, which indicates that the model encodes useful information at the phoneme level.

Lastly, we looked in more depth at which linguistic and acoustic features influence word recognition performance. In general, words that are spoken more slowly have a higher word recognition score. The effect of frequency of a word in the training set on word recognition performance demonstrates how reliant such a model is on its training data. Furthermore, the size of the word-initial cohort was found to have a significant effect on word recognition performance. This shows that, similar to human speech processing, the number of words that match the input speech influence recognition accuracy. It is well known that in human speech recognition, words can be activated or suppressed by priming effects, thus hindering or aiding in recognition [23]. It would be an interesting direction for future research to see if words preceded by a priming context show the expected effects on word recognition performance.

For future research it would be interesting to look at what word a sequence of phonemes is mapped to when it does not retrieve the correct image. This could give more insight into how phonemes are embedded within the model. Also, it would be interesting to see if there are other linguistic or acoustic factors in addition to those we investigated which affect word recognition performance.

5. References

[1] O. Räsänen and H. Rasilo, "A joint model of word segmentation and meaning acquisition through cross-situational learning." *Psychological review*, vol. 122, no. 4, p. 792, 2015.

[2] M. Tomasello, "The usage-based theory of language acquisition," in *The Cambridge handbook of child language*. Cambridge Univ. Press, 2009, pp. 69–87.

[3] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 237–244.

[4] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing Systems*, 2016, pp. 1858–1866.

[5] G. Chrupala, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proceedings of the 55th of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017, p. 613622.

[6] D. Merkx, S. L. Frank, and M. Ernestus, "Language learning using speech to image retrieval," in *Proceedings of Interspeech 2019. Crossroads of Speech and Language*, 2019.

[7] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic keyword spotting by learning from images and speech," *arXiv preprint arXiv:1710.01949*, 2017.

[8] O. Scharenborg, L. Besacier, A. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stker, P. Godard, M. Mller, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang, and E. Dupoux, "Speech technology for unwritten languages," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 964–975, 2020.

[9] H. Kamper and M. Roth, "Visually grounded cross-lingual keyword spotting in speech," *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018.

[10] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137.

[11] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 649–665.

[12] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 1, p. 8998, Jan. 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2018.2872106

[13] W. N. Havard, J.-P. Chevrot, and L. Besacier, "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, Nov. 2019, pp. 339–348.

[14] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," *arXiv preprint arXiv:1911.09602*, 2019.

[15] G. Chrupała, B. Higy, and A. Alishahi, "Analyzing analytical methods: The case of phonology in neural models of spoken language," *arXiv preprint arXiv:2004.07070*, 2020.

[16] A. Weber and O. Scharenborg, "Models of processing: lexicon," *WIREs Cognitive Science*, pp. 387–401, 2012.

[17] D. Norris, J. M. McQueen, and A. Cutler, "Competition and segmentation in spoken-word recognition." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 21, no. 5, p. 1209, 1995.

[18] W. D. Marslen-Wilson and A. Welsh, "Processing interactions and lexical access during word recognition in continuous speech," *Cognitive psychology*, vol. 10, no. 1, pp. 29–63, 1978.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[21] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[22] M. S. Vitevitch and P. A. Luce, "Phonological neighborhood effects in spoken word perception and production," *Annual Review of Linguistics*, vol. 2, pp. 75–94, 2016.

[23] P. R. Chiappe, M. C. Smith, and D. Besner, "Semantic priming in visual word recognition: Activation blocking and domains of processing," *Psychonomic Bulletin & Review*, vol. 3, no. 2, pp. 249–253, 1996.