Learning to recognise words using visually grounded speech

Scholten, Sebastiaan; Merkx, Danny; Scharenborg, Odette

DOI
10.1109/ISCAS51556.2021.9401692

Publication date
2021

Document Version
Accepted author manuscript

Published in
2021 IEEE International Symposium on Circuits and Systems (ISCAS)

Learning to Recognise Words using Visually Grounded Speech

1st Sebastiaan Scholten
Multimedia Computing Group
Delft University of Technology
Delft, The Netherlands
Scholtensebastiaan@gmail.com

2nd Danny Merkx
Centre for Language Studies
Radboud University
Nijmegen, The Netherlands
D.Merkx@let.ru.nl

3rd Odette Scharenborg
Multimedia Computing Group
Delft University of Technology
Delft, The Netherlands
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**

I. Introduction

Babies initially have little understanding of what is being said around them. It is theorized that repeatedly hearing certain words while observing certain objects around them enables babies to learn a mapping between speech and objects [1]. Repeatedly hearing utterances in the context of some functional consistency, like picking up an object, displays 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].

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]–[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 focused on architectural and training scheme improvements [5], [6], [11]and applications such as semantic keyword spotting [7], [12] and speech-based image retrieval [3], [5], [6], [8]. 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]–[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. Real speech recognition is expected to be more challenging due to more variation in quality, noise and speaking rate than in synthetic speech as can be seen for instance in [5].

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.

II. Methodology

A. Visually Grounded Speech Model

In this paper we use the VGS model implemented in [6], with the exception of using a four instead of a three layer GRU. The model consists of two DNNs: a pretrained image encoder (ResNet-152 [19]) 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. We refer to [6] and the freely available PyTorch implementation for more details on the model and feature preprocessing https://github.com/DannyMerkx/speech2image/tree/Interspeech19.

The model was trained in order for matching image-caption pairs to have a similarity larger by a margin α than for mismatched pairs using 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.

We train the model on Flickr8k [20], a database with 8k images and 5 written captions per image for a total of 40k captions. Spoken versions of the captions are collected and provided by Harwath et al. [3]. We use the data split provided by [10]. Caption-to-image retrieval, a standard evaluation metric for our training task, is measured as Recall@N; the percentage of captions for which the correct image was in the top N retrieved images. Images are retrieved based on the cosine distance to the caption embedding.

B. Experiments

In our first experiment, we test the model on isolated words to investigate how well the model learned to map these words onto their visual referents. The second experiment uses a gating paradigm where we test the model on 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. This also allows us to test which

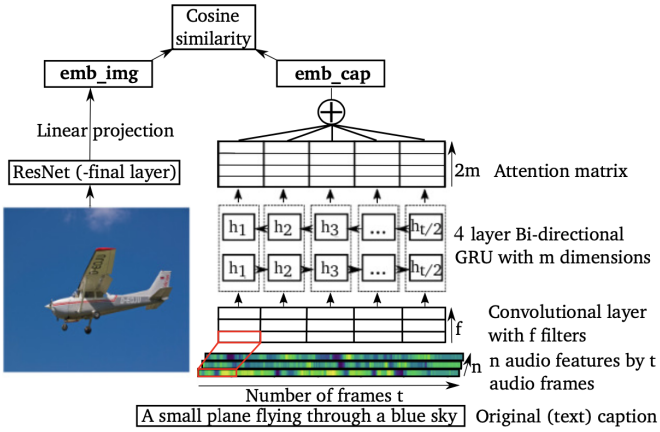

Fig. 1. A visual representation of the image encoder parallel to the caption encoder. Based on [6].

acoustic factors in the speech signal are influencing the model's word recognition performance.

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) 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.

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.

III. Results

The scores in Table I 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.

A. 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.

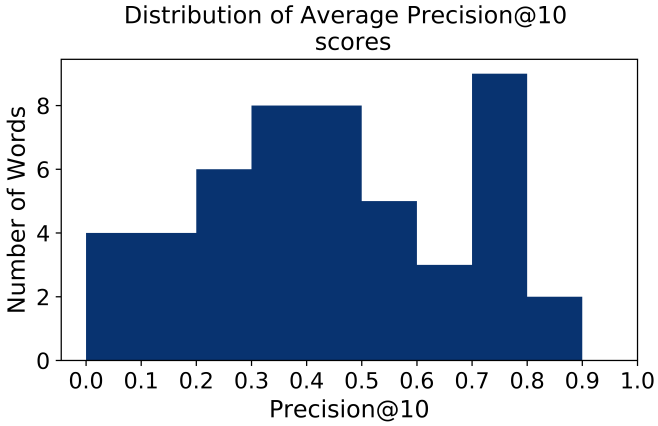

Fig. 2. Distribution of Average P@10 scores for the 49 tested words, assigned to bin intervals of size 0.1.

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

B. 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

TABLE I
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	

TABLE II
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

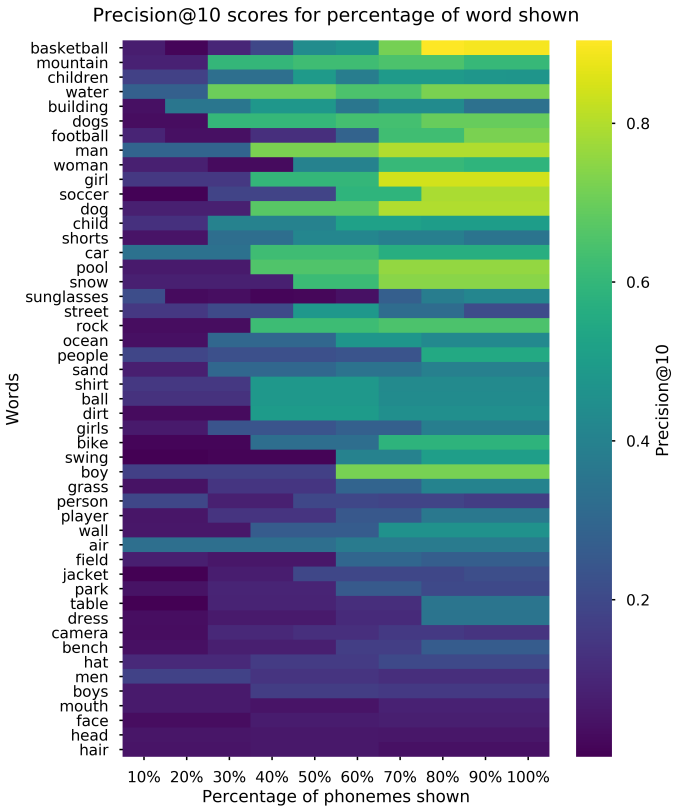

Fig. 3. Heatmap showing the P@10 scores of a given word (shown on the y-axis) as a function of the phoneme sequence length. The x-axis indicates the percentage of phonemes of the word that were available to the model.

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 III. 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 III
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

IV. 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.

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. 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*, 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. 89–98, 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.