Familiar words but strange voices: Modeling the influence of speech variability on word recognition

Alexandra Mayn[1] **Badr M. Abdullah**[2] **Dietrich Klakow**[2]

[1]Department of Information and Computing Sciences
Utrecht University, The Netherlands
[2]Department of Language Science and Technology (LST)
Saarland University, Germany
a.mayn@uu.nl, {babdullah|dietrich}@lsv.uni-saarland.de

Abstract

We present a deep neural model of spoken word recognition which is trained to retrieve the meaning of a word (in the form of a word embedding) given its spoken form, a task which resembles that faced by a human listener. Furthermore, we investigate the influence of variability in speech signals on the model's performance. To this end, we conduct a set of controlled experiments using word-aligned read speech data in German. Our experiments show that (1) the model is more sensitive to dialectical variation than gender variation, and (2) recognition performance of word cognates from related languages reflect the degree of relatedness between languages in our study. Our work highlights the feasibility of modeling human speech perception using deep neural networks.

1 Introduction

Human speech is highly complex and variable. The sources underlying this variability include speaker-related factors such as vocal tract shape, gender, age, and dialect as well as context-related factors such as word surprisal and phonological prominence. As a result, two acoustic realizations of the same word are unlikely to be identical even if produced by the same speaker. Nevertheless, listeners can reliably recognize spoken words despite the lack of acoustic-phonetic invariance in speech (Pisoni and Levi, 2007). The robust human ability to decode the intended message from a highly variable, noisy speech signal enables speakers of different but related languages to communicate with each other using their own mother tongue — a phenomenon that has been referred to as *receptive multilingualism* (Gooskens, 2019).

To gain a better understanding of human speech processing, a vast body of research at the intersection of speech perception and cognitive modeling has been dedicated to developing computational models of spoken word recognition (cf. Weber and Scharenborg (2012) for an overview). In a nutshell, models of spoken word recognition aim to simulate and explain the process of accessing the mental lexicon given a representation of an auditory word stimulus (McClelland and Elman, 1986; Marslen-Wilson, 1987; Norris, 1994; Gaskell and Marslen-Wilson, 1997). Despite the considerable differences in the representational specificity of the proposed models in the literature, there is a consensus among them with respect to the activation of multiple word candidates which leads to competition for lexical access (Weber and Scharenborg, 2012). One model of word recognition that we take inspiration from in this paper is the Distributed Cohort Model (DCM) (Gaskell and Marslen-Wilson, 1997), which is a connectionist model that defines the process of spoken word recognition as a mapping of low-level acoustic features onto the stored semantic and phonological representations, allowing efficient lexical access. A computational model of spoken word recognition allows researchers to simulate the conditions of behavioral experiments on human listeners and investigate whether the predictions of the model show human-like behavior.

Although deep neural networks (DNNs) have become the dominant paradigm for automatic speech recognition (ASR) research in the last decade (Graves et al., 2006; Mohamed et al., 2009; Hinton et al., 2012), using DNN-based ASR components to model human speech processing has only been explored recently with the EARSHOT model (Magnuson et al., 2020). EARSHOT is an incremental model based a long short-term memory (LSTM) that captures the temporal structure of speech. The training data for the EARSHOT model are spoken words produced using a speech synthesizer and each word is associated into a sparse vector that represents the word semantics. The authors

Proceedings of the 16th Conference of the European Chapter of the Associationfor Computational Linguistics: Student Research Workshop, pages 96–102
April 19 - 23, 2021.

use a unique but arbitrary sparse vector for each word, thus the semantic relatedness of words is not encoded in their representations. EARSHOT is trained to map each acoustic word form onto its semantic vector.

In this paper, we attempt to bridge between the connectionist view of word recognition and the recent advances in spoken language learning using deep neural networks. We also address some of the modeling limitations in the EARSHOT model. Precisely, our contribution is two-fold: (1) we propose a model of spoken word recognition based on a deep neural network that maps a spoken word form onto a distributed meaning representation. Our model is trained on naturalistic data that consists of actual acoustic realizations of spoken words extracted from the German portion of the Spoken Wikipedia Corpus. And (2) we investigate the degree to which the emergent representations from the model can generalize with respect to two sources of variability in speech signals — inter-speaker variability and cross-lingual variability.

2 A neural model of spoken word recognition

Our proposed model can be described at the high level as a function that maps the acoustic form of a word onto its lexical meaning. In the following, we describe the different representation schemes of our model.

2.1 Acoustic form representation

Human speech is modeled with various low-level signal representations. In this paper, we adopt the conventional approach in automatic speech recognition (ASR) which converts a time-domain speech waveform into a time-frequency frame-based representation using a standard signal processing pipeline. In particular, we convert each acoustic segment of a spoken word into a sequence of MFCC vectors $\mathbf{x}_{1:T} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_t \in \mathbb{R}^k$ is a spectral feature vector, or a frame, at timestep t and T is the number of frames.

2.2 Meaning representation

Following previous studies that adopted the distributional approach to represent lexical meaning (Pimentel et al., 2019; Williams et al., 2020), we use pre-trained distributed word representations, or word embeddings, as a proxy for the stored lexical representations of word forms. This modeling

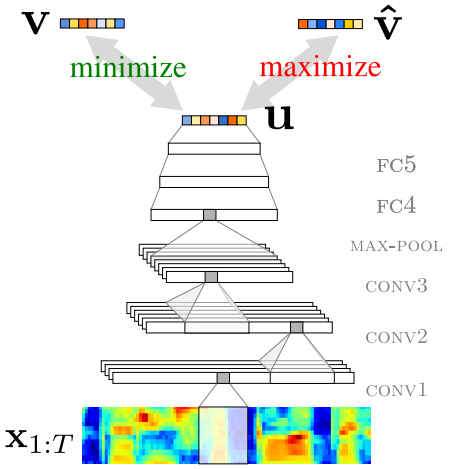

Figure 1: A schematic view of our proposed model for spoken word recognition.

choice can be justified since word embeddings have been shown to reliably encode lexical features such as taxonomic information (Rubinstein et al., 2015).

2.3 Proposed model

Architecture. Similar to the architecture presented in work of Maas et al. (2012), our proposed spoken word recognition model is based on a multi-layer convolutional neural network that maps an acoustic input onto a meaning representation (depicted in Figure 1). However, instead of vector regression as the objective function, the training procedure of our model builds on the ideas of visually-grounded learning of spoken language (Harwath et al., 2016; Chrupała et al., 2017a). While in previous work models have been trained to project an image and its corresponding spoken caption onto a shared representation space, we train our model to project an acoustic segment of a word onto its word embedding. This process can be formalized as a mapping function using a deep neural network as follows:

$$\mathbf{u} = \boldsymbol{f}(\mathbf{x}_{1:T};\ \boldsymbol{\theta})$$

where $\mathbf{u}$ is the meaning representation computed by the model, $\boldsymbol{f}(.)$ is the model presented as a parametric function, $\mathbf{x}_{1:T}$ is the observed acoustic word segment, and $\boldsymbol{\theta}$ are the model's parameters learned in a supervised approach.

Training. Given a training dataset of N tuples $\{(\mathbf{x}^1_{1:T}, \mathbf{v}^1), (\mathbf{x}^2_{1:T}, \mathbf{v}^2), \ldots, (\mathbf{x}^N_{1:T}, \mathbf{v}^N)\}$, our model is trained to take an acoustic word token $\mathbf{x}_{1:T}$ as input, build up a meaning representation $\mathbf{u}$, and then minimize the distance between the computed representation $\mathbf{u}$ and the embedding of the

word $\mathbf{v}$. This learning objective can be realized by projecting the acoustic word token into the word embedding space in such a way that an acoustic segment and embedding of the same word type are encouraged to end up closer in space than mismatched word embeddings. Concretely, we use a triplet margin loss function as follows:

$$\mathcal{L} = \sum_{i=1}^{N} \max\big(0, \alpha + d(\mathbf{u}^i, \mathbf{v}^i) - d(\hat{\mathbf{u}}^i, \mathbf{v}^i)\big) + \max\big(0, \alpha + d(\mathbf{u}^i, \mathbf{v}^i) - d(\mathbf{u}^i, \hat{\mathbf{v}}^i)\big)$$

where $d(.)$ is the cosine distance metric and $\mathbf{u}^i$ and $\mathbf{v}^i$ are the matching computed representation and embedding of a word, while $\hat{\mathbf{u}}^i$ and $\hat{\mathbf{v}}^i$ are the unmatched computed representation and embedding that are sampled from the mini-batch of N samples. α is the margin hyperparameter of the loss function.

3 Experimental setup

3.1 Experimental data

We use the multilingual Spoken Wikipedia Corpus (SWC), which consists of recordings of Wikipedia articles read by volunteers in German, Dutch, and English (Köhn et al., 2016). A large portion of the dataset has been word-aligned and each article is associated with a metadata file that optionally includes (self-identified) information about the speaker's gender and dialect. Therefore, this resource is highly suitable for our experimental aims concerning speech variability.

3.2 Model hyperparameters

Low-level speech features. We use 39-dimensional MFCC feature vectors as well as frame-level averaged energy as low-level speech features. Frames are extracted from speech segment of 25ms with 10ms overlap between frames. Each speech sample is then scaled with word-level zero mean and unit variance.

Speech encoder. We employ three convolutional layers over the temporal dimension with 128, 128, and 256 channels respectively and strides of 1 step for each layer. Batch normalization and ReLU non-linearity are applied after each convolutional operation. The speech representation is down-sampled by applying a single max pooling operation at the end of the convolution block. Then, the resulting vector from the convolutional layers is fed into two fully-connected layers with dropout ($p = 0.5$) and ReLU non-linearity, followed by a linear projection that outputs a representation in the same dimensionality as the word embedding.

	dim	R@1	R@5	R@10
GloVe	300	0.159	0.451	0.608
FastText (FT)	300	0.176	0.461	0.610
Flair	4096	0.216	0.530	0.665
FT + Flair	4396	**0.227**	**0.557**	**0.696**

Table 1: Comparison of the model's retrieval performance using different word embeddings.

Training details. The triplet margin loss is used with $\alpha = 0.2$ for all presented experiments. We use the Adam optimizer with a learning rate of 1×10^{-3} and train our models with a batch size of 32 samples for 60 epochs.

4 Experiments

We present and discuss the results of our experiments in this section. We first investigate the effect of different pre-trained word embeddings on the model performance. In the variability experiments, we aim to probe the model's ability to generalize to unheard speaker types as well as to recognize spoken word cognates in two languages that are phylogenetically related to German: Dutch and English. Following Chrupała et al. (2017b), we use the R @ N metric to evaluate our models for $N = \{1, 5, 10\}$.

4.1 Choice of word embeddings

In this experiment, we train our model on a subset of the SWC consisting of 1500 word types, 20 tokens per type, with each of the following word embeddings: GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017), bidirectional Flair (Akbik et al., 2018), and stacked Flair and FastText.

The retrieval scores of the model with different word embeddings are reported in Table 1. Although the difference is not dramatic, the best-performing model is the one that uses stacked FastText and Flair embeddings. It seems that stacking the embeddings provides richer semantic representations that benefit the model during training. Therefore, we proceed to the variability experiments with stacked Flair and FastText word embeddings as meaning representations.

	R@1	R@5	R@10	med. R
heard	0.473	0.850	0.970	2
speaker	0.348	0.688	0.812	3
gender	0.297	0.615	0.756	3
dialect	0.240	0.515	0.643	5

Table 2: Retrieval performance across speaker types.

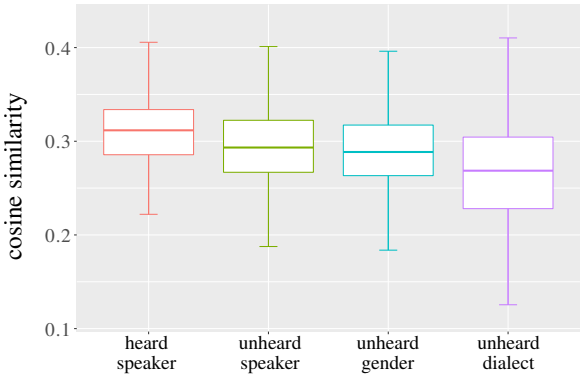

Figure 2: Average cosine similarity of utterance-embedding pairs by speaker type.

4.2 Speaker variability

In this experiment, we aim to probe the model's robustness against speaker variability by comparing its performance on various speaker groups: unheard utterances by heard speakers, unheard speakers, unheard gender (female speakers), and unheard dialect (speakers who self-identified as native speakers of the Swiss German variety). To this end, we train a separate model on a subset of the German portion of the SWC consisting of 2500 word types, 10-100 tokens per type, which were produced by native male speakers of standard German. This training set size is chosen as a trade-off between having a representative training set that includes a variety of words with different lexical properties and practical considerations such as training time and scalability of the model. The test sets we use for evaluation are matched at the token level, the only difference being the speaker characteristics.

Retrieval scores, including median rank of the correct embeddings, are reported in Table 2, and average cosine similarity of the computed meaning representation from the input signal to the corresponding embedding is displayed in Figure 2. Overall, one can observe that signal variability due to speaker-related factors that are unobserved during training degrades the model's performance. A one-way ANOVA test on the cosine similarities revealed significant differences between speaker types ($\chi^2(3) = 230.2, p < 0.001$).[1] Post-hoc Tukey HSD revealed significant differences between all groups except *unheard speaker* and *unheard gender* ($p > 0.5$).

The model performs best at recognizing words when they are spoken by a speaker heard during training, suggesting that the representations learned by the model are not entirely speaker invariant. Interestingly, the model is quite good at generalizing to unheard gender, performing on a par with unheard speakers of the same gender. We hypothesize that the model learns to abstract from pitch variations because there was pitch variability present in the training data. Finally, spoken words from an unheard dialect (i.e., Swiss German) are more challenging for the model to correctly recognize, which suggests that the representations induced by the model are more sensitive to fine-grained acoustic-phonetic variations in the signal than pitch variations.

4.3 Cross-lingual variability

Speakers of related languages are often able to decode some information from each other's speech without ever having had to explicitly learn the correspondences because related languages exhibit prelexical as well as lexical similarities. Gooskens et al. (2018) have shown that the comprehension ability of speakers of related languages correlates very strongly with the degree of language relatedness from a phylogenetic point of view. In this experiment, we explore whether and to what extent the model which has only been exposed to German will be able to recognize cognates in two related languages, English and Dutch. We also ask the question: does the cross-lingual performance reflect the degree of language relatedness? Since German and Dutch are a part of the continental West Germanic dialect continuum, while English is not, we hypothesize that the model should be better at recognizing spoken Dutch words than spoken English words.

To this end, we use the same model as for the speaker variability experiment. The test sets contain cognates in German, English and Dutch, aligned at the token level. Words in the German and Dutch test sets are produced by unheard male speakers of the standard language variety, while

[1] We used cosine similarities for statistical testing because ranks are not normally distributed; most words have low rank.

	R@1	R@5	R@10	med. R
German	0.388	0.715	0.819	2
Dutch	0.041	0.138	0.203	133.5
English	0.011	0.064	0.111	177.5
chance	≈ 0.0004	≈ 0.002	≈ 0.04	—

Table 3: Retrieval performance on cognates.

words in the English test set are produced by male native speakers of American English.[2] Spoken word representations for Dutch and English are obtained via a forward pass through the speech encoder, the same way as for German, and the model receives no explicit information that the cognates are in a different language.

Retrieval scores @1, 5 and 10, as well as median rank, for all three languages are reported in Table 3. Average cosine similarities of matching utterance-embedding pairs for the three languages are reported in Figure 3. The standard error is relatively high for the two related languages, especially for Dutch, because the model's guess for some cognates was quite poor. One-way repeated measures ANOVA reveals unsurprising significant differences between groups ($\chi^2(2)=362.3$, $p<0.0001$): the model is much better at recognizing German since this is the language that the model was trained on. If we compare the retrieval scores for Dutch and English to chance performance,[3] we observe that the model is relatively good at recognizing cognates in the two related languages, with 20% and 11% of words in Dutch and English respectively within the top 10 retrieved word embeddings. The difference in performance on the two related languages is shown to be significant in post-hoc Tukey HSD ($p=0.004$), supporting our hypothesis that cross-lingual word recognition performance reflects language relatedness.

We look more closely at the model's recognition performance on cognates (a selection is reported in Figure 4). Cognates which are well-recognized are mostly identical word forms, except for vowel length or slightly different consonant quality (e.g., the Dutch *jaar* (/jaːr/) and the German *Jahr* (/jaːʁ/). However, other interesting correspondences are apparent. For example, we observe that for the word *ship* (/ʃɪp/), the model is better at recognizing the English word, which is different from the German *Schiff* (/ʃɪf/) only in the final consonant, than the Dutch *schip* (/sxɪp/), where the word onset is different. This finding suggests that the model might have learned to pay closer attention to the beginnings of words. Future work could explore systematically which sound correspondences make it easy or difficult to recognize cognates.

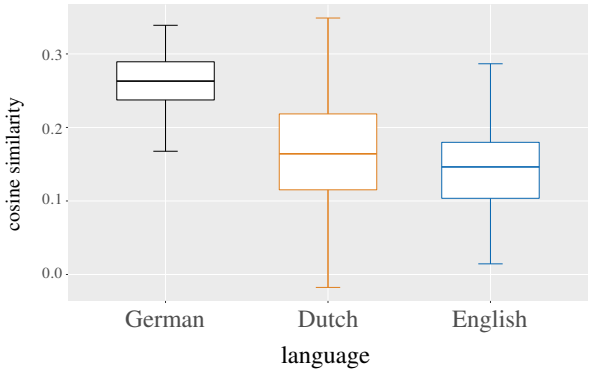

Figure 3: Average cosine similarity of utterance-embedding pairs by language.

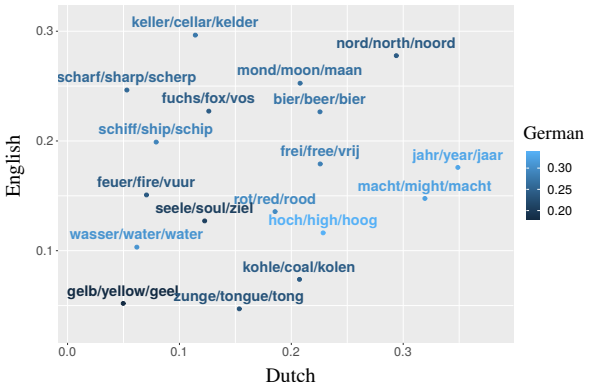

Figure 4: Average cosine similarity between matching utterance-pair embeddings for cognates in the three languages (words depicted as German/English/Dutch). Higher cosine similarity corresponds to a more accurate acoustic representation and, in turn, better recognition. For example, the word *cellar* is recognized better in English than in Dutch (reflected in a relatively higher cosine similarity). Lighter shades of blue correspond to higher cosine similarity for German utterances.

[2] We would have favoured to include British English in the study as well but that was not feasible since not enough data of that kind is available in the SWC.

[3] We approximate chance performance by assuming that the probability of a word ending up at each of the 2500 positions is equally high. This approximation is not perfect since it does not take into account the fact that more frequent words are relatively more likely to end up in the top positions. However, this is very computationally expensive to calculate.

5 Discussion and future work

We observe that the representations produced by the model seem to be largely gender-invariant since the model's performance on unheard female speakers is on a par with its performance on unheard male speakers. On the other hand, dialect variability seems to have a stronger impact than gender which suggests that the model is sensitive to low-level acoustic-phonetic variance in the speech signal. We would expect a human listener to exhibit similar patterns in case of little exposure to dialectical variability.

Our model operates by creating a general representation of a word, which it uses to generalize to unheard speakers. However, there is evidence in psycholinguistics which suggests that we adapt to individuals' pronunciation and create speaker-specific representations (Kleinschmidt and Jaeger, 2015). This could be simulated by fine-tuning the trained model on more data by a particular speaker.

When tested on cognates in related languages in a zero-shot fashion, the model shows reasonably good cognate recognition performance. There is also a significant difference in the model's performance on Dutch and on English, reflecting the closer phylogenetic relationship between German and Dutch. One could imagine using the proposed model to test mutual intelligibility: if trained on Dutch, would such a model be better at recognizing German cognates than the other way around? This would be a test of intelligibility that eliminates extra-linguistic factors that cannot be isolated in behavioral experiments (van Heuven et al., 2012).

Since this is a word-level model of word recognition, there is no facilitatory effect of context, which human listeners are known to rely on to a large extent when there is uncertainty as to which word was uttered. In the cross-lingual experiment, too, we would expect that a model which is able to benefit from context would show much better performance. Such sentence-level models of related language comprehension are an exciting avenue to pursue in future work.

Acknowledgments

We would like to thank the anonymous reviewers of the student research workshop at EACL for their comments and suggestions. Badr M. Abdullah is supported by funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project ID 232722074, SFB 1102.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017a. Representations of language in a model of visually grounded speech signal. *arXiv preprint arXiv:1702.01991*.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017b. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622, Vancouver, Canada. Association for Computational Linguistics.

Gareth M Gaskell and William D Marslen-Wilson. 1997. Integrating form and meaning: A distributed model of speech perception. *Language and cognitive Processes*, 12(5-6):613–656.

Charlotte Gooskens. 2019. 8 receptive multilingualism. *Multidisciplinary Perspectives on Multilingualism: The Fundamentals*, 19:149.

Charlotte Gooskens, Vincent J van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte, and Stefanie Voigt. 2018. Mutual intelligibility between closely related languages in europe. *International Journal of Multilingualism*, 15(2):169–193.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

David Harwath, Antonio Torralba, and James Glass. 2016. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866.

Vincent J van Heuven, Charlotte Gooskens, and Renée van Bezooijen. 2012. Mutual intelligibility of dutch and german cognates.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Dave F Kleinschmidt and Florian T Jaeger. 2015. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, 122(2):148.

Arne Köhn, Florian Stegen, and Timo Baumann. 2016. Mining the spoken wikipedia for speech data and beyond.

Andrew L Maas, Stephen D Miller, Tyler M O'neil, Andrew Y Ng, and Patrick Nguyen. 2012. Word-level acoustic modeling with convolutional vector regression. In *Proc. ICML Workshop Representation Learn*.

James S Magnuson, Heejo You, Sahil Luthra, Monica Li, Hosung Nam, Monty Escabi, Kevin Brown, Paul D Allopenna, Rachel M Theodore, Nicholas Monto, et al. 2020. Earshot: A minimal neural network model of incremental human speech recognition. *Cognitive science*, 44(4):e12823.

William Marslen-Wilson. 1987. Functional parallelism in spoken word-recognition. *Cognition*, 25:71–102.

James L McClelland and Jeffrey L Elman. 1986. The TRACE model of speech perception. *Cognitive Psychology*, 18(1):1 – 86.

Abdel-rahman Mohamed, George Dahl, and Geoffrey Hinton. 2009. Deep belief networks for phone recognition. In *NIPS workshop on deep learning for speech recognition and related applications*, volume 1, page 39. Vancouver, Canada.

Dennis Norris. 1994. Shortlist: a connectionist model of continuous speech recognition. *Cognition*, 52(3):189 – 234.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Tiago Pimentel, Arya D McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764.

David B Pisoni and Susannah Levi. 2007. Some observations on representations and representational specificity in speech perception and spoken word recognition. In *The Oxford Handbook of Psycholinguistics*, pages 3–18. Oxford University Press.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? *Volume 2: Short Papers*.

Andrea Weber and Odette Scharenborg. 2012. Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3):387–401.

Adina Williams, Tiago Pimentel, Arya McCarthy, Hagen Blix, Eleanor Chodroff, and Ryan Cotterell. 2020. Predicting declension class from form and meaning. In *Proceedings of the 58th Annual Meeting for the Association of Computational Linguistics*. York.