On the difficulty of a distributional semantics of spoken language

Grzegorz Chrupała
Tilburg University
g.chrupala@uvt.nl

Lieke Gelderloos
Tilburg University
l.j.gelderloos@uvt.nl

Ákos Kádár
Tilburg University
a.kadar@uvt.nl

Afra Alishahi
Tilburg University
a.alishahi@uvt.nl

Abstract

In the domain of unsupervised learning most work on speech has focused on discovering low-level constructs such as phoneme inventories or word-like units. In contrast, for written language, where there is a large body of work on unsupervised induction of semantic representations of words, whole sentences and longer texts. In this study we examine the challenges of adapting these approaches from written to spoken language. We conjecture that unsupervised learning of the semantics of spoken language becomes feasible if we abstract from the surface variability. We simulate this setting with a dataset of utterances spoken by a realistic but uniform synthetic voice. We evaluate two simple unsupervised models which, to varying degrees of success, learn semantic representations of speech fragments. Finally we present inconclusive results on human speech, and discuss the challenges inherent in learning distributional semantic representations on unrestricted natural spoken language.

1 Introduction

In the realm of NLP for written language, unsupervised approaches to inducing semantic representations of words have a long pedigree and a history of substantial success (Landauer et al., 1998; Blei et al., 2003; Mikolov et al., 2013b). The core idea behind these models is to build word representations that can predict their surrounding context. In search for similarly generic and versatile representations of whole sentences, various composition operators have been applied on word representations (e.g. Socher et al., 2013; Kalchbrenner et al., 2014; Kim, 2014; Zhao et al., 2015). Alternatively, sentence representations are induced via the objective to predict the surrounding sentences (e.g. Le and Mikolov, 2014; Kiros et al., 2015; Arora et al., 2016; Jernite et al., 2017; Logeswaran and Lee, 2018). Such representations capture aspects of the meaning of the encoded sentences, which can be used in a variety of tasks such as semantic entailment or text understanding.

In the case of spoken language, unsupervised methods usually focus on discovering relatively low-level constructs such as phoneme inventories or word-like units. This is mainly due to the fact that the key insight from distributional semantics that “you shall know the word by the company it keeps” (Firth, 1957) is hopelessly confounded in the case of spoken language. In text two words are considered semantically similar if they co-occur with similar neighbors. However, speech segments which occur in the same utterance or situation often have many other features in addition to similar meaning, such as being uttered by the same speaker or accompanied by similar ambient noise.

In this study we show that if we can abstract away from speaker and background noise, we can effectively capture semantic characteristics of spoken utterances in an unsupervised way. We present SegMatch, a model trained to match segments of the same utterance. SegMatch utterance encodings are compared to those in Audio2Vec, which is trained to decode the context that surrounds an utterance. To investigate whether our representations capture semantics, we evaluate on speech and vision datasets where photographic images are paired with spoken descriptions. Our experiments show that for a single synthetic voice, a simple model trained only on image captions can capture pairwise similarities that correlate with those in the visual space.

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 167-173.
New York City, New York, January 3-6, 2019

Furthermore we discuss the factors preventing effective learning in datasets with multiple human speakers: these include confounds between semantic and situational factors as well as artifacts in the datasets.

2 Related work

Studies of unsupervised learning from speech typically aim to discover the phonemic or lexical building blocks of the language signal. Park and Glass (2008) show that words and phrase units in continuous speech can be discovered using algorithms based on dynamic time warping. van den Oord et al. (2017) introduce a Vector Quantised-Variational AutoEncoder model, in which a convolutional encoder trained on raw audio data gives discrete encodings that are closely related to phonemes. Recently several unsupervised speech recognition methods were proposed that segment speech and cluster the resulting word-like segments (Kamper et al., 2017a) or encode them into segment embeddings containing phonetic information (Wang et al., 2018). Scharenborg et al. (2018) show that word and phrase units arise as a by-product in end-to-end tasks such as speech-to-speech translation. In the current work, the aim is to directly extract semantic, rather than word form information from speech.

Semantic information encoded in speech is used in studies that ground speech to the visual context. Datasets of images paired with spoken captions can be used to train multimodal models that extract visually salient semantic information from speech, without access to textual information (Harwath and Glass, 2015; Harwath et al., 2016; Kamper et al., 2017b; Chrupała et al., 2017; Alishahi et al., 2017; Harwath and Glass, 2017). This form of semantic supervision, through contextual information from another modality, has its limits: it can only help to learn to understand speech describing the here and now.

On the other hand, the success of word embeddings derived by distributional semantic principles has shown how rich the semantic information within the structure of language itself is. Semantic representations of words obtained through Latent Semantic Analysis have proven to closely resemble human semantic knowledge (Blei et al., 2003; Landauer et al., 1998). Word2vec models produce semantically rich word embeddings by learning to predict the surrounding words in text (Mikolov et al., 2013a,b) and this principle is extended to sentences in the Skip-thought model (Kiros et al., 2015) and several subsequent works (Arora et al., 2016; Jernite et al., 2017; Logeswaran and Lee, 2018).

In the realm of spoken language, in Chung and Glass (2017) the sequence-to-sequence Audio2vec model learns semantic embeddings for audio segments corresponding to words, by predicting the audio segments around it. Chung and Glass (2018) further experiment with this model and rename it to Speech2vec. Chen et al. (2018) train semantic word embeddings from word-segmented speech as part of their method of training an ASR system from non-aligned speech and text. These works are closely related to our current study, but crucially, unlike them we do *not* assume that speech is already segmented into discrete words.

3 Models

3.1 Encoder

All the models in this section use the same encoder architecture. The encoder is loosely based on the architecture of Chrupała et al. (2017), i.e. it consists of a 1-dimensional convolutional layer which subsamples the input, followed by a stack of recurrent layers, followed by a self-attention operator. Unlike Chrupała et al. (2017) we use GRU layers (Chung et al., 2014) instead of RHN layers (Zilly et al., 2017), and do not implement residual connections. These modifications are made in order to exploit the fast native CUDNN implementation of a GRU stack and thus speed up experimentation in this exploratory stage of our research. The encoder Enc is defined as follows:

$$\mathrm{Enc}(\mathbf{x}) = \mathrm{unit}(\mathrm{Attn}(\mathrm{GRU}_\ell(\mathrm{Conv}_{\mathrm{s,d,z}}(\mathbf{x})))) \tag{1}$$

where Conv is a convolutional layer with length s, d channels, and stride z, GRU_ℓ is a stack of ℓ GRU layers, Attn is self-attention and unit is L2-normalization. The self-attention operator computes a weighted sum of the RNN activations at all timesteps:

$$\mathrm{Attn}(\mathbf{x}) = \sum_t \alpha_t \mathbf{x}_t \tag{2}$$

where the weights α_t are determined by an MLP with learned parameters $\mathbf{U}$ and $\mathbf{W}$, and passed through the timewise softmax function:

$$\alpha_t = \frac{\exp(\mathbf{U}\tanh(\mathbf{W}\mathbf{x}_t))}{\sum_{t'}\exp(\mathbf{U}\tanh(\mathbf{W}\mathbf{x}_{t'}))} \tag{3}$$

3.2 Audio2vec

Firstly we define a model inspired by Chung and Glass (2017) which uses the multilayer GRU encoder described above, and a single-layer GRU decoder, conditioned on the output of the encoder.

The model of Chung and Glass (2017) works on word-segmented speech: the encoder encodes the middle word of a five word sequence, and the decoder decodes each of the surrounding words. Similarly, the Skip-thought model of (Kiros et al., 2015) works with a sequence of three sentences, encoding the middle one and decoding the previous and next one. In our fully unsupervised setup we do not have access to word segmentation, and thus our Audio2vec models work with arbitrary speech segments. We split each utterance into three equal sized chunks: the model encodes the middle one, and decodes the first and third one.

The decoder predicts the MFCC features at time $t+1$ based on the state of the hidden layer at time t. From reading Chung and Glass (2017) it is not clear whether in addition to the hidden state their decoder also receives the MFCC frame at t as input. We thus implemented two versions, one with and one without this input.

Audio2vec-C The decoder receives the output of the encoder as the initial state of the hidden layer, and the frame at t as input as it predicts the next frame at $t+1$.

$$\hat{\mathbf{x}}^{\text{first}}_{t+1} = \mathbf{F}\mathbf{h}_t \tag{4}$$

$$\mathbf{h}_t = \text{gru}(\mathbf{h}_{t-1}, \mathbf{x}^{\text{first}}_t) \tag{5}$$

$$\mathbf{h}_0 = \text{Enc}\left(\mathbf{x}^{\text{middle}}\right) \tag{6}$$

where $\mathbf{x}^{\text{first}}_t$ are the MFCC features of the previous chunk at time t, $\hat{\mathbf{x}}^{\text{first}}_{t+1}$ are the predicted features at the next time step, $\mathbf{F}$ is a learned projection matrix, $\text{gru}(\cdot,\cdot)$ is a single step of the GRU recurrence, and $\mathbf{x}^{\text{middle}}$ is the sequence of the MFCC features of the input. The decoder for the third chunk $\mathbf{x}^{\text{third}}$ is defined in the same way.

Audio2vec-U The decoder receives the output of the encoder as the input at each time step, but does not have access to the frame at t.

$$\hat{\mathbf{x}}^{\text{first}}_{t+1} = \mathbf{F}\mathbf{h}_t \tag{7}$$

$$\mathbf{h}_t = \text{gru}(\mathbf{h}_{t-1}, \text{Enc}(\mathbf{x}^{\text{middle}})) \tag{8}$$

In this version $\mathbf{h}_0$ is a learned parameter. There are two separate decoders: i.e. the weights of the decoder for the first chunk and for the third chunk are *not* shared.

For both versions of **Audio2vec** the loss function is the Mean Squared Error.

3.3 SegMatch

This model works with segments of utterances also: we split each utterance approximately in half, while erasing a short portion in the center in order to prevent the model from finding trivial solutions based on matching local patterns at the edges of the segments. The encoder is as described above. After encoding the segments, we project the initial and final segments via separate learned projection matrices:

$$\mathbf{b} = \mathbf{B}\text{Enc}(\mathbf{x}_{0:m}) \tag{9}$$

$$\mathbf{e} = \mathbf{E}\text{Enc}(\mathbf{x}_{m+k:n}) \tag{10}$$

where $\mathbf{x}_{0:n}$ is the sequence of MFCC frames for an utterance, k is the size of the erased segment, $\text{Enc}(\cdot)$ is the encoder and $\mathbf{B}$ and $\mathbf{E}$ are the projection matrices for the beginning and end segment respectively. That is, there is a single shared encoder for both types of speech segments (beginning and end), but the projections are separate. There is no decoding, but rather the model learns to match encoded segments from the same utterance and distinguish them from encoded segments from different utterances within the same minibatch. The loss function is similar to the one for matching spoken utterances to images in Chrupała et al. (2017), with the difference that here we are matching utterance segments to each other:

$$\mathcal{L} = \sum_{\mathbf{b},\mathbf{e}}\left(\sum_{\mathbf{b}'}\max[0, \alpha + d(\mathbf{b},\mathbf{e}) - d(\mathbf{b}',\mathbf{e})] + \sum_{\mathbf{e}'}\max[0, \alpha + d(\mathbf{b},\mathbf{e}) - d(\mathbf{b},\mathbf{e}')]\right) \tag{11}$$

where $(\mathbf{b},\mathbf{e})$ are beginning and end segments from the same utterance, and $(\mathbf{b}',\mathbf{e})$ and $(\mathbf{b},\mathbf{e}')$ are beginning and end segments from two different utterances within a batch, while $d(\cdot,\cdot)$ is the cosine distance between encoded segments. The loss function thus attempts to make the cosine distance between encodings of matching segments less than the distance between encodings of mismatching segment pairs, by a margin.

Note that the specific way we segment speech is not a crucial component of either of the models: it is mostly driven by the fact that we run our experiments on speech and vision datasets, where speech consists of isolated utterances. For data consisting of longer narratives, or dialogs, we could use different segmentation schemes.

4 Experimental setup

4.1 Datasets

In order to facilitate evaluation of the semantic aspect of the learned representations, we work with speech and vision datasets, which couple photographic images with their spoken descriptions. Thanks to the structure of these data we can use the evaluation metrics detailed in section 4.2.

Synthetically spoken COCO This dataset was created by Chrupała et al. (2017), based on the original COCO dataset (Lin et al., 2014), using the Google TTS API. The captions are spoken by a single synthetic voice, which is realistic but simpler than human speakers, lacking variability and ambient noise. There are 300,000 images, each with five captions. Five thousand images each are held out for validation and test.

Flickr8k Audio Caption Corpus This dataset (Harwath and Glass, 2015) contains the captions in the original Flickr8K corpus (Hodosh et al., 2013) read aloud by crowdworkers. There are 8,000 images, each image with five descriptions. One thousand images are held out for validation, and another one thousand for the test set.

Places Audio Caption Corpus This dataset was collected by (Harwath et al., 2016) using crowdworkers. Here each image is described by a single spontaneously spoken caption. There are 214,585 training images, and 1000 validation images (there are no separate test data).

4.2 Evaluation metrics

We evaluate the quality of the learned semantic speech representations according to the following criteria.

Paraphrase retrieval For the Synthetically Spoken COCO dataset as well as for the Flickr8k Audio Caption Corpus each image is described via five independent spoken captions. Thus captions describing the same image are effectively paraphrases of each other. This structure of the data allows us to use a paraphrasing retrieval task as a measure of the semantic quality of the learned speech embeddings. We encode each of the spoken utterances in the validation data, and rank the others according to the cosine similarity. We then measure: (a) Median rank of the top-ranked paraphrase; and (b) recall@K: the proportion of paraphrases among K top-ranked utterances, for $K \in \{1, 5, 10\}$.

Representational similarity to image space Representational similarity analysis (RSA) is a way of evaluating how pairwise similarities between objects are correlated in two object representation spaces (Kriegeskorte et al., 2008). Here we compare cosine similarities among encoded utterances versus cosine similarities among vector representations of images. Specifically, we create two pairwise $N \times N$ similarity matrices: (a) among encoded utterances from the validation data, and (b) among images corresponding to each utterance in (a). Note that since there are five descriptions per image, each image is replicated five times in matrix (b). We then take the upper triangulars of these matrices (excluding the diagonal) and compute Pearson's correlation coefficient between them. The image features for this evaluation are obtained from the final fully connected layer of VGG-16 (Simonyan and Zisserman, 2014) pre-trained on Imagenet (Russakovsky et al., 2014) and consist of 4096 dimensions.

4.3 Settings

We preprocess the audio by extracting 12-dimensional mel-frequency cepstral coefficients (MFCC) plus log of the total energy. We use 25 milisecond windows, sampled every 10 miliseconds. Audio2vec and SegMatch models are trained for a maximum of 15 epochs with Adam, with learning rate 0.0002, and gradient clipping at 2.0. SegMatch uses margin $\alpha = 0.2$. The encoder GRU has 5 layers of 512 units. The convolutional layer has 64 channels, size of 6 and stride 3. The hidden layer of the attention MLP is 512. The GRU of the Audio2vec decoder has 512 hidden units; the size of the output of the projections $\mathbf{B}$ and $\mathbf{E}$ in SegMatch is also 512 units. For SegMatch the size of the erased center portion of the utterance is 30 frames. We apply early stopping and report all the results of each model after the epoch for which it scored best on recall@10. When applying SegMatch on human data, each mini-batch includes utterances spoken only by one speaker: this is in order to discourage the model from encoding speaker-specific features.

5 Results

5.1 Synthetic speech

Table 1 shows the evaluation results on synthetic speech. Representations learned by Audio2vec and SegMatch are compared to the performance of random vectors, mean MFCC vectors, as well as visually supervised representations (VGS, model from Chrupała et al. (2017)). Audio2vec works better than chance and mean MFCC on paraphrase retrieval, but does not correlate with the visual space. SegMatch works much better than Audio2vec according to both criteria. It does not come close to VGS on paraphrase retrieval, but it does correlate with the visual modality even better.

5.2 Human speech

Places This dataset only features a single caption per image and thus we only evaluate according to RSA: with both SegMatch and Audio2vec we found the correlations to be zero.

Flickr8K Initial experiments with Flickr8K were similarly unsuccessful. Analysis of the learned SegMatch representations revealed that in spite of partitioning the data by speaker for training, speaker identity can be decoded from them.

Enforcing speaker invariance We thus implemented a version of SegMatch where an auxiliary speaker classifier is connected to the encoder via a gradient reversal operator (Ganin and Lempitsky, 2015). This architecture optimizes the main loss, while at the same time pushing the encoder to remove information about speaker identity from the representation it outputs. In preliminary experiments we saw that this addition was able to prevent speaker identity from being encoded in the representations during the first few epochs of training. Evaluating this speaker-invariant representation gave contradictory results, shown in Table 2: very good scores on paraphrase retrieval, but zero correlation with visual space.

Further analysis showed that there seems to be an artifact in the Flickr8K data where spoken captions belonging to consecutively numbered images share some characteristics, even though the images do not. As a side effect, this causes captions belonging to the same image to also share some features, independent of their semantic content, leading to high paraphrasing scores. The artifact may be due to changes in data collection procedure which affected some aspect of the captions in ways which correlate with their sequential ordering in the dataset.

If we treat the image ID number as a regression target, and the first two principal components of the SegMatch representation of one of its captions as the predictors, we can account for about 12% of the holdout variance in IDs using a non-linear model (using either K-Nearest Neighbors or Random Forest). This effect disappears if we arbitrarily relabel images.

6 Conclusion

For synthetic speech the SegMatch approach to inducing utterance embeddings shows very promising performance. Likewise, previous work has shown some success with word-segmented speech. There remain challenges in carrying over these results to natural, unsegmented speech. Word segmentation is a highly non-trivial research problem in itself and the variability of spoken language is a serious and intractable confounding factor.

Even when controlling for speaker identity there are still superficial features of the speech signal which make it easy for the model to ignore the semantic content. Some of these may be due to artifacts in datasets and thus care is needed when evaluating unsupervised models of spoken language: for example use of multiple evaluation criteria may help spot spurious results. In spite of these challenges, in future we want to further explore the effectiveness of enforcing desired invariances via auxiliary classifiers with gradient reversal.

References

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

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

	Recall@10 (%)	Median rank	$\mathrm{RSA}_{\mathrm{image}}$
VGS	27	6	0.4
SegMatch	**10**	**37**	**0.5**
Audio2vec-U	5	105	0.0
Audio2vec-C	2	647	0.0
Mean MFCC	1	1,414	0.0
Chance	0	3,955	0.0

Table 1: Results on Synthetically Spoken COCO. The row labeled VGS is the visually supervised model from Chrupała et al. (2017).

	Recall@10 (%)	Median rank	$\mathrm{RSA}_{\mathrm{image}}$
VGS	15	17	0.2
SegMatch	12	17	0.0
Mean MFCC	0	711	0.0

Table 2: Results on Flickr8K. The row labeled VGS is the visually supervised model from Chrupała et al. (2017).

Yi-Chen Chen, Chia-Hao Shen, Sung-Feng Huang, and Hung-yi Lee. 2018. Towards unsupervised automatic speech recognition trained by unaligned speech and text only. *arXiv preprint arXiv:1803.10952*.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Deep Learning and Representation Learning Workshop*.

Yu-An Chung and James Glass. 2017. Learning word embeddings from speech. In *NIPS ML4Audio Workshop*.

Yu-An Chung and James Glass. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930-1955, volume 1952-59. The Philological Society.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *IEEE Automatic Speech Recognition and Understanding Workshop*.

David Harwath and James R Glass. 2017. Learning word-like units from joint audio-visual analysis. *arXiv preprint arXiv:1701.07481*.

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.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Yacine Jernite, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2017a. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46:154–174.

Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. 2017b. Visually grounded learning of keyword prediction from untranscribed speech. In *Proc. Interspeech 2017*, pages 3677–3681.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3276–3284.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *CoRR*, abs/1711.00937.

Alex S Park and James R Glass. 2008. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):186–197.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. ImageNet large scale visual recognition challenge.

Odette Scharenborg, Laurent Besacier, Alan Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, et al. 2018. Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the Speaking Rosetta JSALT 2017 workshop. *arXiv preprint arXiv:1802.05092*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Yu-Hsuan Wang, Hung-yi Lee, and Lin-shan Lee. 2018. Segmental audio word2vec: Representing utterances as sequences of vectors with applications in spoken term detection. In *2018 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings.*

Han Zhao, Zhengdong Lu, and Pascal Poupart. 2015. Self-adaptive hierarchical sentence model. In *IJCAI*, pages 4069–4076.

Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. 2017. Recurrent highway networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4189–4198, International Convention Centre, Sydney, Australia. PMLR.