Towards Localisation of Keywords in Speech Using Weak Supervision

Kayode Olaleye* **Benjamin van Niekerk** **Herman Kamper**
Department of E&E Engineering
Stellenbosch University

Abstract

Developments in weakly supervised and self-supervised models could enable speech technology in low-resource settings where full transcriptions are not available. We consider whether keyword localisation is possible using two forms of weak supervision where location information is not provided explicitly. In the first, only the presence or absence of a word is indicated, i.e. a bag-of-words (BoW) labelling. In the second, visual context is provided in the form of an image paired with an unlabelled utterance; a model then needs to be trained in a self-supervised fashion using the paired data. For keyword localisation, we adapt a saliency-based method typically used in the vision domain. We compare this to an existing technique that performs localisation as a part of the network architecture. While the saliency-based method is more flexible (it can be applied without architectural restrictions), we identify a critical limitation when using it for keyword localisation. Of the two forms of supervision, the visually trained model performs worse than the BoW-trained model. We show qualitatively that the visually trained model sometimes locate semantically related words, but this is not consistent. While our results show that there is some signal allowing for localisation, it also calls for other localisation methods better matched to these forms of weak supervision.

1 Introduction

There is a growing body of work considering how speech processing systems can be developed in the absence of conventional transcriptions [30, 8, 22, 27, 32]. Several of these studies consider the setting where we have an indication of whether a word occurs in an utterance or not, but don't know where the word occurs. Given this weak form of supervision, we ask whether it is still possible to localise words in a speech utterance.

We specifically consider two types of weak supervision. In the first, speech audio is paired with a bag-of-words (BoW) labelling, indicating the presence or absence of a word without giving the location, order, or number of occurrences [22]. This is useful when only noisy labels are available, e.g., for low-resource languages. In the second, images are paired with spoken captions. This form of visual supervision can be useful when it is not possible to collect textual labels, e.g., for languages without a written form. Since both the utterance and the paired image are unlabelled, some form of self-supervision is required, where a proxy task is used to train the model [7, 1, 20, 21, 31]. Existing approaches include using an external image tagger to extract soft multi-class labels [23, 16], or a model can be trained so that the images and the speech are projected into a joint embedding space [14, 4]. Compared to full transcriptions, both forms of supervision are closer to the signals that infants would have access to while learning their first language [24, 25, 10, 9, 2, 3], and to how one would teach new words to robots using spoken language [29]. Moreover, these weak forms

*Correspondence: kaykola.olaleye@gmail.com

34th Self-Supervised Learning for Speech and Audio Processing Workshop @ NeurIPS 2020.

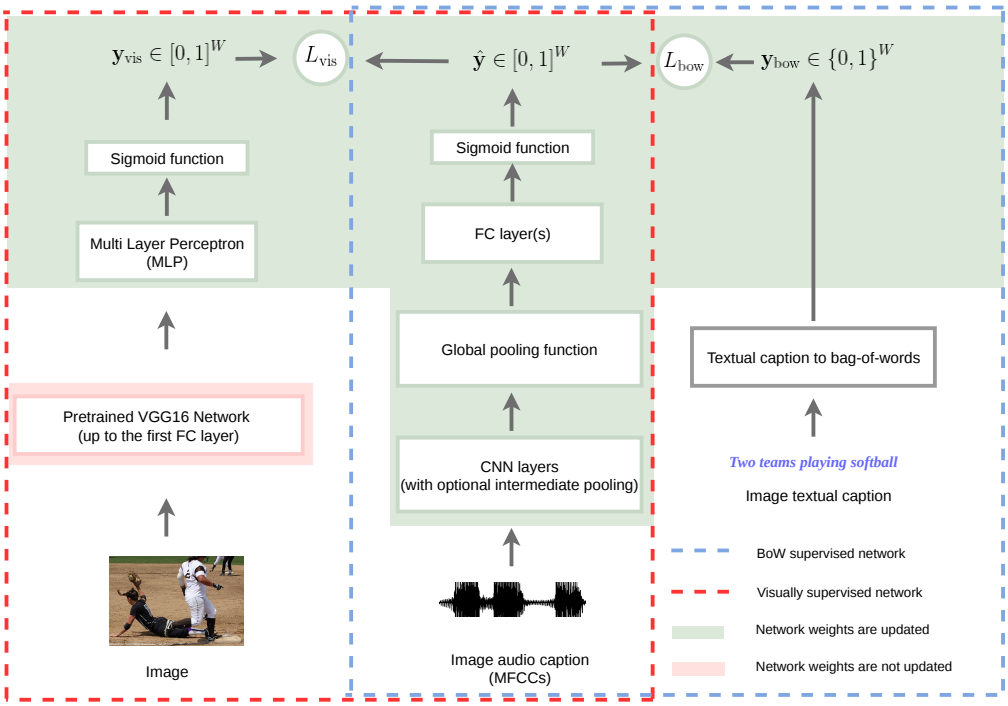

Figure 1: We consider localisation with networks trained with bag-of-word (blue, right) and visual (red, left) supervision.

of supervision could conceivably be easier to obtain when developing systems for low-resource languages [5].

We consider two localisation mechanisms. The first was introduced in [22], referred to as PSC based on the author names. PSC uses a convolutional neural network (CNN) architecture to jointly locate and classify words using BoW supervision. Here we extend PSC by also considering visual supervision. The second is GradCAM [26], a saliency-based method originally built for locating objects in images using the gradients of a target concept with respect to filter activations. Here we apply GradCAM for localisation of keywords in speech with both forms of supervision.

We find that PSC-based localisation outperforms GradCAM. However, the underlying model used by GradCAM performs better on word *detection* (the task of identifying whether a word is present in an utterance or not without considering location). We speculate that this is due to a mismatch between the multi-label classification loss used here and the activation estimation method of GradCAM, which was originally developed for single-label multi-class classification. Unsurprisingly, we find that BoW-trained models outperform visually trained models on the localisation task. We show qualitatively that this is sometimes caused by the visual supervision capturing semantic and other information from the scene not matching the spoken caption exactly. However, although this aligns with previous studies on other tasks [4, 17], this finding is not always consistent. Taken together, our results suggest that other self-supervision and localisation methods need to be considered that are better matched to the form of multi-label weak supervision used here.

2 Proposed Localisation Mechanisms and Models

Two forms of weak supervision. In the absence of full transcriptions, Palaz et al. [22] considered a weak form of supervision where each speech sequence $X = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$ is paired with BoW labels $\mathbf{y}_{\text{bow}} \in \{0,1\}^W$ (blue dotted region on the right in Figure 1). Each element $y_{\text{bow},w}$ is an indicator for whether word w occurs in the utterance. The utterance X consists of T acoustic vectors (Mel-frequency cepstral coefficients in our case), and W is the total number of words in the vocabulary.

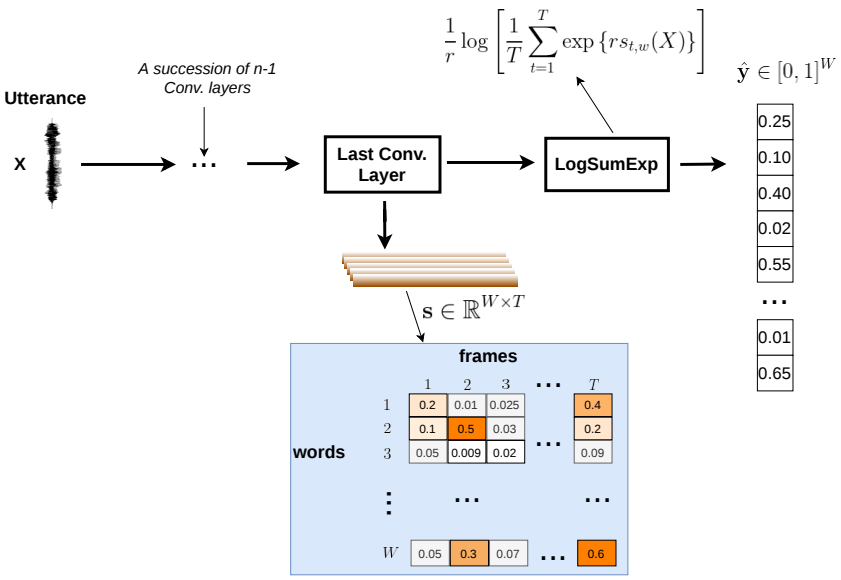

Figure 2: An illustration of the PSC model. The blue region represents how the output of the last convolutional layer is unpacked to give a score $s_{t,w}$ for each frame t and each word w. In this example, the proposed location for the keywords corresponding to indexes 2 and W would respectively be frames 2 and T, which achieve the highest scores in the depicted blue region.

In another form of weak supervision, each utterance X is paired with a corresponding image I (red region, Figure 1 left). In our case I is a scene and X is a spoken caption [12, 14, 13]. Kamper et al. [17] proposed that, in an attempt to get the same type of supervision as in the BoW case, I can be passed through a trained multi-label image tagger. The tagger is trained on external data, and can therefore be seen as a way to utilise existing vision systems to obtain a noisy target which can be used for self-supervision. Concretely, the tagger outputs soft probabilities $\mathbf{y}_{\text{vis}} \in [0,1]^W$ for whether a particular word is relevant to the image. We use the same visual tagger as in [17].

The PSC model. The PSC model is designed to simultaneously perform detection and localisation of keywords in speech utterances when provided with BoW targets [22]. The structure of the model is illustrated in Figure 2. The model consists of a number of convolutional layers which operates on X to produce a score for the presence or absence of a word at a particular time-step. Concretely, the model's last one-dimensional convolutional layer produces filter outputs $\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_T$, where each $\mathbf{s}_t \in \mathbb{R}^W$ is a W-dimensional vector. We therefore have one convolutional filter per word, with $s_{t,w}$ giving the score for word w at time t (depicted as the blue region in Figure 2). These frame-level localisation scores are fed into an aggregation function to produce a single utterance-level detection score:

$$g_w(X) = \frac{1}{r}\log\left[\frac{1}{T}\sum_{t=1}^{T}\exp\{rs_{t,w}(X)\}\right] \tag{1}$$

This LogSumExp aggregation function is equivalent to average pooling when $r \to 0$ and max pooling when $r \to \infty$. According to Palaz et al. [22], this intermediate aggregation operation drives the weights of frames which have similar scores close to each other during training, resulting in better localisation performance. The final output of the network is the probability that each word is present in the utterance: $\hat{\mathbf{y}} = \sigma(\mathbf{s}(X))$, with σ the sigmoid function. The model is trained with the summed binary log loss:

$$L(\hat{\mathbf{y}}, \mathbf{y}_{\text{bow}}) = -\sum_{w=1}^{W}\{y_{\text{bow},w}\log\hat{y}_w + (1-y_{\text{bow},w})\log[1-\log\hat{y}_w]\} \tag{2}$$

where $\mathbf{y}_{\text{bow}}$ indicates that BoW supervision is used. For visual supervision, it is replaced by $\mathbf{y}_{\text{vis}}$. Note that in this model, the localisation mechanism is built into the model by directly connecting the

last convolutional layer with the set of words. If we had any other layers between the scoring layer and the output, this connection would be lost.

The GradCAM model. In contrast to PSC, GradCAM is a method that can be used for localisation in any CNN architecture [26]. However, here we apply it within a particular model. We therefore refer to both the localisation mechanism and our specific CNN architecture together as the "GradCAM model" (although these are technically disjoint). Our GradCAM model again has a number of convolutional layers, but also uses intermediate max-pooling layers, max-pooling over time over its last convolutional layer, and a number of fully connected layers before terminating in its output (see Section 3). The output is again denoted as $\hat{\mathbf{y}} \in [0,1]^W$ and the same loss (2) as for PSC is used. After training the model, GradCAM localisation is performed as follows. Let us call the output from the last one-dimensional convolutional layer $\mathbf{q}$, i.e., it produces $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_T$, where each $\mathbf{q}_t \in \mathbb{R}^K$ is a K-dimensional vector with K the number of filters. (In the PSC model, the number of filters in the last convolutional layer had to be W, but we do not have this constraint here.) We first determine the "importance" of the k^{th} filter to the word w:

$$\gamma_{k,w} = \frac{1}{T} \sum_{t=1}^{T} \frac{\partial \hat{y}_w}{\partial q_{t,k}} \tag{3}$$

This value indicates how closely the k^{th} convolutional filter pays attention to word w, based on the trained model weights. For a particular utterance, we want to know the scores $s_{t,w}$ of word w at time t. We obtain this by calculating the values of all the filters, and then weighing each filter by its importance $\gamma_{k,w}$. We are only interested in changes that would result in a higher output score for a word, so we take the ReLU, resulting in $s_{t,w} = \text{ReLU}\left[\sum_{k=1}^{K} \gamma_{k,w} q_{t,k}\right]$.

3 Experimental Setup and Evaluation

We use the Flickr8k Audio Caption Corpus of Harwath and Glass [12], consisting of five English audio captions for each of the roughly 8k images. 30k utterances are used for training, 5k for development, and 5k for testing. To obtain BoW labels for the training data, we construct indicator vectors $\mathbf{y}_{\text{bow}} \in \{0,1\}^W$ for the $W = 1000$ most common words in the transcriptions. For visual supervision, each training image is passed through the multi-label visual tagger of Kamper et al. [17], which uses VGG-16 [28] and is trained on images [6, 19, 11] disjoint from the data considered here. The result is soft labels $\mathbf{y}_{\text{vis}} \in [0,1]^W$ for the $W = 1000$ image classes from the tagger.

Our PSC model consists of six one-dimensional convolutional layers with ReLU activations. The first has 96 filters with a kernel width of 9 frames. The next four has a width of 11 units. The last convolutional layer, with $W = 1000$ filters and a width of 11 units, is fed into the LogSumExp aggregation function with a final sigmoid activation. Our GradCAM model consists of three one-dimensional convolutional layers with ReLU activations. Intermediate max pooling over 3 units are applied in the first two layers. The first convolution has 64 filters with a width of 9 frames. The second layer has 256 filters with a width of 11 units, and the last layer has 1024 filters with a width of 11. Global max pooling is applied followed by a sigmoid activation to obtain the final output for the $W = 1000$ words. All models are implemented in PyTorch and uses Adam optimisation [18] with a learning rate of $1 \cdot 10^{-4}$.

As in [22], we consider two settings for evaluating localisation performance. In the *oracle* setting, we assume that the system perfectly detects whether a word occurs in an utterance or not. We then evaluate whether it is also able to locate it. The position of the highest score is taken as the predicted location: $\tau_p = \text{argmax}_t(s_{t,w})$. τ_p is accepted as correct if the frame τ_p is within the interval corresponding to the true w, according to forced alignments. Figure 3 depicts two localisation attempts by a model. The proposed location for the keyword *snow* is τ_p and its ground truth location spans frames 160 and 200. In (**a**), τ_p is accepted as a correct location of *snow* because it is within the ground truth frames, and rejected in (**b**) because it is outside the ground truth frames. In the *actual* evaluation setting, detection is also taken into account. Before evaluating localisation, we apply a threshold α to the detection score for w. We then compute the precision, recall, $F1$, and accuracy scores, again using τ_p as above.

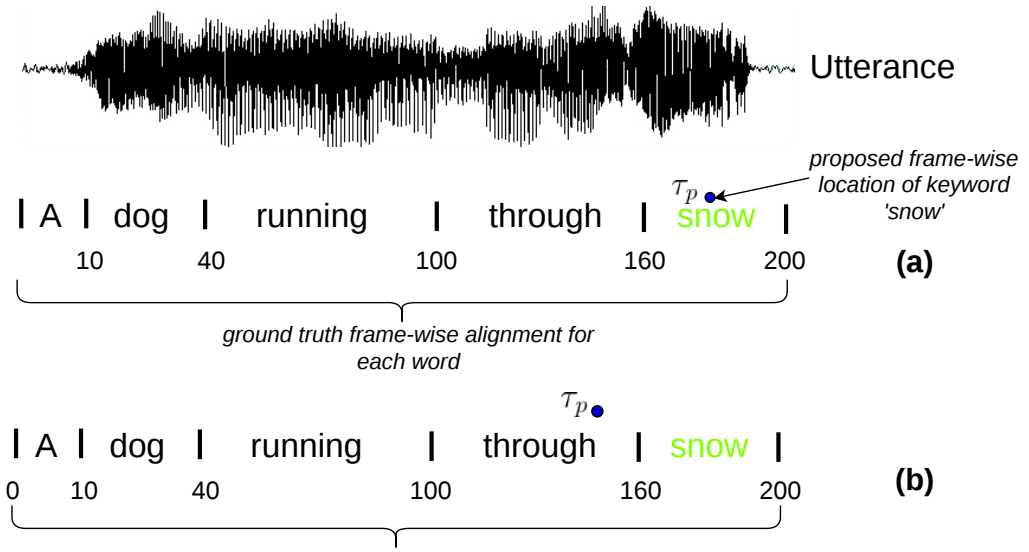

Figure 3: Illustration of the evaluation procedure: (**a**) depicts a successful localisation attempt of the keyword in green (*snow*), while (**b**) depicts a failed attempt.

4 Results and Discussion

Table 1 shows the localisation accuracies in the *oracle* setting for both the PSC and GradCAM models when supervised with BoW labels and visual context. Table 2 gives the scores in the *actual* setting, where detection is also taken into account, using a threshold of $\alpha = 0.4$.

Of the two forms of supervision, BoW labels leads to consistently better localisation than visual supervision. This is not surprising since different speakers could describe the same image in many different ways. Moreover, the visual tagger (which provides the training signal here) can assign high probabilities to concepts that no speaker would refer to (but which is nevertheless present in an image) or could tag semantically related words. As qualitative evidence of the last-mentioned, Figure 4(b) shows that "escalator" is localised when prompted with "stairs". Despite many incorrect predictions when using visual supervision, e.g. Figure 4(c), visual supervision does provide some signal for localisation, with the PSC model achieving a precision of almost 30% in Table 2. Figure 4(a) shows a

	Supervision method	
Mechanism	BoW	Visual
PSC	63.6	19.1
GradCAM	17.8	16.0

Table 1: Oracle localisation accuracy (%) when assuming perfect detection.

	BoW				Visual			
Mechanism	P	R	$F1$	Accuracy	P	R	$F1$	Accuracy
PSC	75.2	53.0	62.2	50.4	28.6	8.0	12.5	7.6
GradCAM	17.7	24.5	20.5	13.2	5.0	5.7	5.3	4.4

Table 2: Actual localisation precision, recall, $F1$ and accuracy (%) when taking detection into account with a threshold of $\alpha = 0.4$.

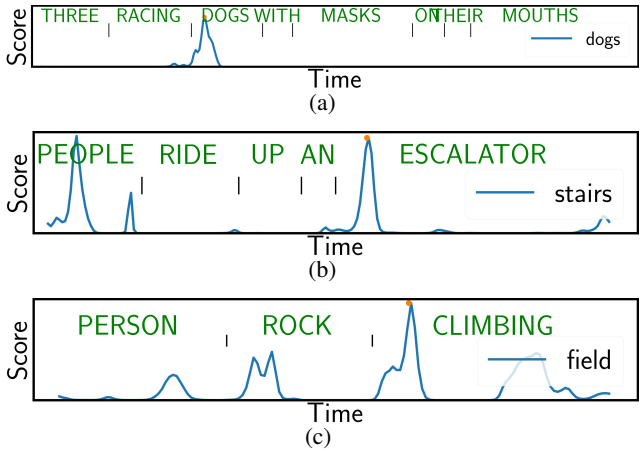

Figure 4: Examples of localisation with the visually supervised PSC model. The keyword being localised is shown on the right of each plot.

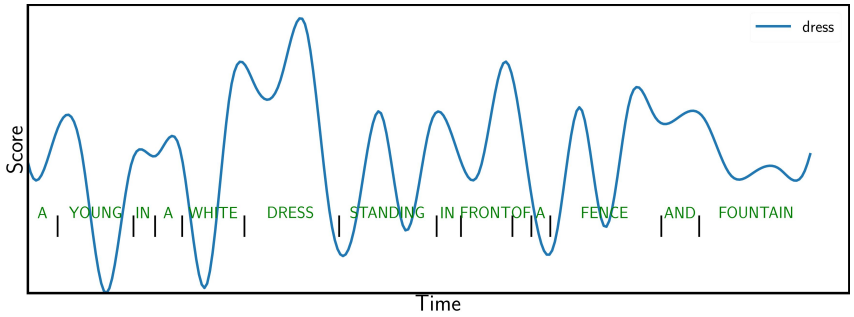

Figure 5: An example localisation with the GradCAM model for the keyword "dress".

correct localisation. This is noteworthy since this model is trained without any textual labels and is still able to make text predictions.

Of the two localisation mechanisms, PSC outperforms GradCAM on all metrics with both forms of supervision. Figure 5 gives an example of GradCAM localisation. We see that, in contrast to PSC (Figure 4), peaks are produced on many words.We speculate that this is because GradCAM tries to identify the parts of the input that will cause a large change in a particular output unit. In single-label multi-class classification, for which GradCAM was developed, a higher probability for a particular output implies lower probabilities for others. But this is not the case for multi-label classification, as used here, and the gradients of multiple words could therefore affect the output, and this would be captured in the gradients. Table 3 shows that when considering detection, i.e. only predicting whether a word is present in an utterance without evaluating localisation, the GradCAM model in fact outperforms the PSC model.

5 Conclusion

We asked whether keyword localisation in speech is possible with two forms of weak supervision when location information is not provided. We compared two localisation methods with two forms of supervision: bag-of-word (BoW) labels and visual context. While a saliency-based method performed poorly, a method where localisation is performed as part of the network performed well with BoW supervision and showed that visual supervision does provide signal for higher precision localisation. As far as we know, this is the first work to use these localisation mechanisms with visual supervision. Harwath et al. [15] did consider localisation of a word given a corresponding image, but this is different from our study where we locate a word in speech based on a given text label.

	$\alpha = 0.4$			$\alpha = 0.6$		
Model	P	R	$F1$	P	R	$F1$
Visual supervision:						
PSC	44.5	9.8	16.1	74.7	4.3	8.1
GradCAM	29.3	22.0	25.1	42.7	12.7	19.6
BoW supervision:						
PSC	82.2	49.0	61.4	87.8	46.1	60.4
GradCAM	79.3	52.6	63.2	82.5	50.9	63.0

Table 3: Keyword detection scores (without considering localisation) with threshold α.

Our results suggests a mismatch between saliency-based localisation and the multi-label model used here, with a superior detection model performing poorly in localisation. This suggest that better localisation should be possible given a mechanism better aligned to the model and multi-label classification loss.

Acknowledgements

This work is supported in part by the National Research Foundation of South Africa (grant number: 120409), a Google Faculty Award for the last author, and Google Africa PhD scholarships for the first two authors.

References

[1] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. *In Proc. NeurIPS*, 2016.

[2] P. C. Bomba and E. R. Siqueland. The nature and structure of infant form categories. *Journal of Experimental Child Psychology*, 1983.

[3] L. Boves, L. ten Bosch, and R. Moore. Acorns-towards computational modeling of communication and recognition skills. In *In Proc. ICCI*, 2007.

[4] G. Chrupała, L. Gelderloos, and A. Alishahi. Representations of language in a model of visually grounded speech signal. *In Proc. ACL*, 2017.

[5] V. De Sa and D. Ballard. Category learning through multimodality sensing. *Neural Comput.*, 1998.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *In Proc. CVPR*, 2009.

[7] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. *In Proc. ICCV*, 2017.

[8] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn. An attentional model for speech translation without transcription. *In Proc. NAACL HLT*, 2016.

[9] P. D. Eimas and P. C. Quinn. Studies on the formation of perceptually based basic-level categories in young infants. *Child Development*, 1994.

[10] L. Gelderloos and G. Chrupała. From phonemes to images: levels of representation in a recurrent neural model of visually-grounded language learning". *In Proc. COLING*, 2016.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *In Proc. CVPR*, 2014.

[12] D. Harwath and J. Glass. Deep multimodal semantic embeddings for speech and images. *In Proc. ASRU*, 2015.

[13] D. Harwath and J. Glass. Learning word-like units from joint audio-visual analysis. *In Proc. ACL*, 2017.

[14] D. Harwath, A. Torralba, and J. Glass. Unsupervised learning of spoken language with visual context. *In Proc. NeurIPS*, 2016.

[15] 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 Proc. ECCV*, 2018.

[16] H. Kamper, A. Anastassiou, and K. Livescu. Semantic query-by-example speech search using visual grounding. *In Proc. ICASSP*, 2019.

[17] H. Kamper, G. Shakhnarovich, and K. Livescu. Semantic speech retrieval with a visually grounded model of untranscribed speech. *IEEE/ACM Trans. Acoust., Speech, Signal Process.*, 2019.

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

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *In Proc. NeurIPS*, 2012.

[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. *In Proc. ICML*, 2011.

[21] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. *In Proc. ECCV*, 2016.

[22] D. Palaz, G. Synnaeve, and R. Collobert. Jointly learning to locate and classify words using convolutional networks. *In Proc. Interspeech*, 2016.

[23] A. Pasad, B. Shi, H. Kamper, and K. Livescu. On the contributions of visual and textual supervision in low-resource semantic speech retrieval. *In Proc. Interspeech*, 2019.

[24] S. Pinker. The language instinct. *Harper Perennial, New York*, 1994.

[25] D. Roy. Grounded spoken language acquisition: experiments in word learning. *IEEE Trans. Multimedia*, 2003.

[26] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *In Proc. ICCV*, 2017.

[27] S. Settle, K. Levin, H. Kamper, and K. Livescu. Query-by-example search with discriminative neural acoustic word embeddings. *In Proc. Interspeech*, 2017.

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

[29] M. Sun and H. Van hamme. Joint training of non-negative tucker decomposition and discrete density hidden markov models. *Comput. Speech Lang.*, 2013.

[30] G. Synnaeve, M. Versteegh, and E. Dupoux. Learning words from images and speech. *NeurIPS Workshop Learn*, 2014.

[31] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. *In Proc. ICCV*, 2015.

[32] R. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen. Sequence-to-sequence models can directly translate foreign speech. *In Proc. Interspeech*, 2017.