INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Visually grounded learning of keyword prediction from untranscribed speech

Herman Kamper, Shane Settle, Gregory Shakhnarovich, Karen Livescu

Toyota Technological Institute at Chicago
{kamperh, settle.shane, greg, klivescu}@ttic.edu

Abstract

During language acquisition, infants have the benefit of visual cues to ground spoken language. Robots similarly have access to audio and visual sensors. Recent work has shown that images and spoken captions can be mapped into a meaningful common space, allowing images to be retrieved using speech and vice versa. In this setting of images paired with untranscribed spoken captions, we consider whether computer vision systems can be used to obtain textual labels for the speech. Concretely, we use an image-to-words multi-label visual classifier to tag images with soft textual labels, and then train a neural network to map from the speech to these soft targets. We show that the resulting speech system is able to predict which words occur in an utterance—acting as a spoken bag-of-words classifier—without seeing any parallel speech and text. We find that the model often confuses semantically related words, e.g. "man" and "person", making it even more effective as a *semantic* keyword spotter.
Index Terms: multimodal modelling, visual semantics, keyword spotting, word discovery, language acquisition

1. Introduction

Current automatic speech recognition (ASR) systems use supervised models trained on huge amounts of annotated resources. In an effort to alleviate this dependence on labelled data, there is growing interest in methods that can learn from untranscribed speech [1–5]. Here we consider the problem of grounding unlabelled speech when paired with images. Annotating speech is expensive and sometimes impossible, e.g. for endangered or unwritten languages [6]; grounding speech using co-occurring visual contexts could be a way to train systems in such low-resource scenarios [7]. This setting is also relevant in robotics, where audio and visual signals can be combined for learning new commands [8–10], and for understanding language acquisition in humans, who have access to visual cues for grounding [11–14].

Specifically, we are interested in the setting considered in [15, 16], where natural images of scenes are paired with spoken descriptions, and neither the images nor speech are labelled. Both [15] and [16] used paired neural networks to map images and speech into a common semantic space where matched images and spoken captions are close to each other. This approach allows images to be retrieved using speech and vice versa. The same task was also considered in earlier work on tagging mobile phone images with spoken descriptions [17, 18]. Despite the practical relevance, and interesting extensions in follow-on work [7, 19], this joint mapping approach does not give an explicit grounding of speech in terms of textual labels.

Here we consider the possibility of using externally trained computer vision systems, which do have access to textual labels, to provide (noisy) supervision for untranscribed speech. Concretely, we use an external image-to-words multi-label visual classifier, predicting for an image a set of words that refer to aspects of the scene. Using soft labels (probabilities) from this vision system, we train a convolutional neural network to map spoken captions to these soft unordered word targets. The result is a speech model that can predict which words (from a fixed vocabulary defined by the vision system) occur in a spoken utterance—acting as a spoken bag-of-words (BoW) classifier.

The previous work in this setting [7, 15, 16, 19] also makes use of intermediate features from pretrained vision models. Our approach can be seen as a further way to exploit vision systems, by also using their textual classification output.

We first apply our word prediction model to two tasks: BoW prediction, where the aim is to predict an unordered set of words that occur in a given utterance, and keyword spotting, where the task is to retrieve all utterances in a collection that contain a given textual keyword. Promising results are achieved on both tasks. Analysis shows that many of the model errors are semantically related to the correct labels, e.g. the model retrieves the speech utterance "a dog runs in the grass" for the textual keyword "field". These "errors" may be desirable in certain settings. So in a final task, we evaluate our model as a *semantic* keyword spotter, where it achieves performance much closer to that of an oracle model trained using ground-truth transcriptions.

2. Related work

Our work intersects with several other research directions. Recent studies have shown that using extra visual features from the scene in which the speech occurs can improve conventional ASR [20, 21]. These systems still rely on labelled speech data, while our aim is to use vision to ground *untranscribed* speech. There has also been much interest in developing speech models that, instead of exact transcriptions, can learn from very noisy labels [22–26]. The study of [26] particularly influenced our approach, since they build a speech system using textual BoW labels (§3.1). In the vision community, image captioning has received much recent attention, where the goal is to produce a fluent and informative natural language description for a visual scene [27–30]. In natural language processing, images have also been used to capture aspects of meaning (semantics) of written language; see [31, 32] for reviews. Other studies have considered multimodal modelling of sounds (not speech) with text and images [33–35], and phonemes with images [36].

3. Word prediction from images and speech

Given a corpus of parallel images and spoken captions, neither with textual labels, we propose a method to train a spoken word prediction model using labels obtained from the visual modality.

3.1. Model overview

Every training image I is paired with a spoken caption of frames $X = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$ (e.g. MFCCs). We use a vision system to tag I with soft textual labels, giving targets to train the speech network $\boldsymbol{f}(X)$ to predict which words are present in X. The network $\boldsymbol{f}(X)$ therefore acts as a spoken bag-of-words (BoW)

http://dx.doi.org/10.21437/Interspeech.2017-502

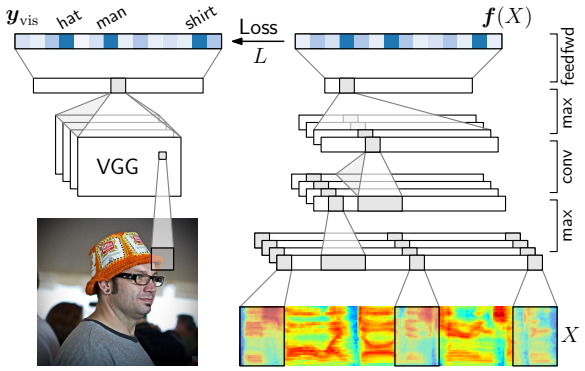

Figure 1: *A multi-label visual classifier is used to produce targets for training a word prediction model using only parallel images and unlabelled spoken captions.*

classifier (disregarding the order and quantity of words). No transcriptions are used during training. When applying the trained $\boldsymbol{f}(X)$, only speech input is used (and no image). The approach is illustrated in Figure 1, and below we give complete details.

If we knew which words occur in training utterance X, we could construct a multi-hot vector $\boldsymbol{y}_{\text{bow}} \in \{0,1\}^W$, with W the vocabulary size, and each dimension $y_{\text{bow},w}$ a binary indicator for whether word w occurs in X. In [26], transcriptions were used to obtain this type of ideal BoW supervision. Instead of a transcription for X, we only have access to the paired image I. We use a multi-label visual classifier (with parameters $\boldsymbol{\gamma}$) which, instead of binary indicators, produces soft targets $\boldsymbol{y}_{\text{vis}} \in [0,1]^W$, with $y_{\text{vis},w} = P(w|I,\boldsymbol{\gamma})$ the probability of word w being present given image I. In Figure 1, $\boldsymbol{y}_{\text{vis}}$ would ideally be close to 1 for w corresponding to words such as "hat", "man" and "shirt", and close to 0 for irrelevant dimensions. This vision system is fixed: during training (below), vision parameters $\boldsymbol{\gamma}$ are never updated.

Given $\boldsymbol{y}_{\text{vis}}$ as target, we train the word prediction model $\boldsymbol{f}(X)$. This model (with parameters $\boldsymbol{\theta}$) consists of a convolutional neural network (CNN) over the speech X, as shown on the right in Figure 1. We interpret each dimension of the output as $f_w(X) = P(w|X,\boldsymbol{\theta})$. Note that $\boldsymbol{f}(X)$ is not a distribution over the vocabulary, since any number of terms in the vocabulary can be present in an utterance; rather, each dimension $f_w(X)$ can have any value in $[0,1]$. We train this speech network using the cross-entropy loss, which (for a single training example) is:

$$L(\boldsymbol{f}(X), \boldsymbol{y}_{\text{vis}}) = -\sum_{w=1}^{W} \{y_{\text{vis},w} \log f_w(X) \; + \; (1 - y_{\text{vis},w}) \log [1 - f_w(X)]\} \quad (1)$$

If we had $y_{\text{vis},w} \in \{0,1\}$, as in $\boldsymbol{y}_{\text{bow}}$, this could be described as the summed log loss of W binary classifiers. The size-W vocabulary of our system is implicitly specified by the vision system.

3.2. Two convolutional architectures over speech

We consider two different convolutional architectures for $\boldsymbol{f}(X)$. Both deal with the variable number of frames in X by pooling over the entire output of their final convolution layer. As input layer, both use a one-dimensional convolution only over time, covering a number of frames and the entire frequency axis.

The first architecture is shown schematically in Figure 1. It is a CNN based on [16, 37], consisting of several convolution and max pooling layers (final pooling covering the entire output), followed by fully connected layers. A sigmoid activation is used for the final output $\boldsymbol{f}(X)$, and ReLUs in intermediate layers.

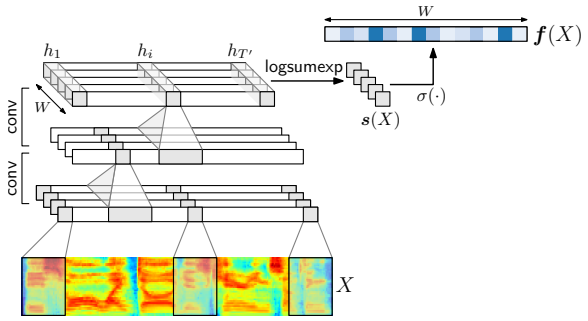

Figure 2: *A two-layer **P**alaz, **S**ynnaeve and **C**ollobert (PSC) network [26]. The rest of our approach is as in Figure 1.*

The second architecture is the one from **P**alaz, **S**ynnaeve and **C**ollobert [26], referred to as PSC. It was originally developed for ideal BoW supervision (§3.1), with the aim of not only doing spoken BoW classification, but also locating where words occur in the speech. PSC aims to do this by explicitly building the vocabulary into its final convolutional layer, as illustrated in Figure 2. The final convolution is linear with W output filters, matching the system vocabulary. The outputs of these final filters are $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{T'}$, with $\boldsymbol{h}_i \in \mathbb{R}^W$. The idea is that $h_{i,w}$ gives a score for word w occurring in the time span corresponding to output i, thus giving an estimate of where w would occur in X. To obtain the network output $\boldsymbol{s}(X) \in \mathbb{R}^W$, PSC does not use mean or max pooling, but rather an intermediate option:

$$s_w(X) = \frac{1}{r} \log \left[\frac{1}{T'} \sum_{t=1}^{T'} \exp\left(r \, h_{t,w}(X)\right) \right] \quad (2)$$

with $s_w(X)$ giving an overall unnormalized score for word w being present in X. This logsumexp pooling is equivalent to mean pooling when $r \to 0$ and max pooling for $r \to \infty$; Palaz *et al.* note that this intermediate method improves PSC's location prediction capability (refer to [26]). The final output of the network is $\boldsymbol{f}(X) = \sigma(\boldsymbol{s}(X))$, with σ the sigmoid function. We use $r = 1$, ReLU activations, and no intermediate pooling.

3.3. The vision system

In image captioning, the goal is to produce a natural language description of a scene [27–30]. In contrast, rather than a fluent sentence, here we want a vision tagging system [38–40] that predicts an unordered set of words (nouns, adjectives, verbs) that accurately describe aspects of the scene (Figure 1, left). This is a multi-label binary classification task, where for each word we must predict whether it is appropriate for the image.

We train our vision tagging system on the Flickr30k data set [41], which contains 30k images, each with a set of 5 captions, which we convert into a BoW after removing stop words. Given a limited set of task-specific training data, such as Flickr30k, a common approach is to start with a visual representation learned as a part of end-to-end training on a larger data set (possibly for a different task), and then adapt it to the task at hand. We follow the established practice of using a representation trained for the ImageNet classification task [42], as also in prior work [7, 16].[1] Specifically, we use VGG-16 [43], but replace the final classification layer with four 3072-unit ReLU layers, followed by a binary classifier for word occurrence. We train this multi-label visual classifier (with parameters $\boldsymbol{\gamma}$) on Flickr30k, with the

[1] The ImageNet output itself is not well-suited to our setting, since it performs a single multi-way classification among a set of image classes.

output layer limited to the $W = 1000$ most common word types in the image captions. The VGG-16 parameters are fixed during training; only the final layers that we add on top are trained.

Note that we train the vision system here only on Flickr30k images that do not correspond to train or test instances from the parallel image-speech data used in our experiments (§4). This leaves around 25k images.[2] Also note that the vision system is trained and then fixed (parameters $\boldsymbol{\gamma}$ is not updated in §4).

4. Experiments

4.1. Experimental setup

We train our word prediction model on the data set of parallel images and spoken captions of [44], containing 8000 images with 5 spoken captions each. The audio comprises around 37 hours of active speech. The data comes with train, development and test splits containing 30 000, 5000 and 5000 utterances, respectively. Speech is parametrized as MFCCs with first and second order derivatives, giving 39-dimensional input.[3] Utterances longer than 8 s are truncated (99.5% of utterances are shorter than 8 s).

Training images are passed through the vision system (§3.3), producing soft targets $\boldsymbol{y}_{\text{vis}}$ for training the word prediction model $\boldsymbol{f}(X)$ on the unlabelled speech. We consider two architectures for $\boldsymbol{f}(X)$, referred to as VisionSpeechCNN and VisionSpeechPSC, respectively (see §3.2). VisionSpeechCNN is structured as follows: 1-D ReLU convolution with 64 filters over 9 frames; max pooling over 3 units; 1-D ReLU convolution with 256 filters over 10 units; max pooling over 3 units; 1-D ReLU convolution with 1024 filters over 11 units; max pooling over all units; 4096-unit fully-connected ReLU; and the 1000-unit sigmoid output. VisionSpeechPSC is structured as follows: 1-D ReLU convolution with 96 filters over 9 frames; four 1-D ReLU convolutions, each with 96 filters over 10 units; 1-D linear convolution with $W = 1000$ filters over 10 units; and logsumexp pooling followed by the final sigmoid activation. We arrived at these two structures starting from those in [16] and [26], respectively, and then tuned them on our development data.

We also obtain upper and lower bounds on performance. As an upper bound, we train two oracle models, OracleSpeechCNN and OracleSpeechPSC, with the same structures as the two VisionSpeech models above. These models are trained on ideal BoW supervision (§3.1): we obtain $\boldsymbol{y}_{\text{bow}}$ targets for the 1000 most common words in the transcriptions of the 30 000 speech training utterances, after removing stop words. Next, as a lower-bound baseline, we use a unigram language model prior that gives the unigram probability of each keyword as estimated from the transcriptions. This baseline gives an indication of how much better our models do than simply hypothesizing common words. Note that the textual transcriptions are used *only* for the baseline and oracle models and for evaluation: neither of the VisionSpeech models ever see any parallel speech and text.

All models were implemented in TensorFlow [45].[4] Based on development tuning, we use Adam optimization [46] with a learning rate of 0.0001 for all models, except those based on PSC, which uses 0.001.

4.2. Spoken bag-of-words prediction

We first consider the task of predicting which words are present in a given test utterance. Given input X, our model gives a score $f_w(X) \in [0, 1]$ for every word w in its vocabulary, and these can be used for spoken BoW prediction. To make a hard prediction, we set a threshold α and output labels for all w where $f_w(X) > \alpha$. We compare the predicted BoW labels to the true set of words in the transcriptions, and calculate precision, recall and F-score across *all* word types in the reference transcriptions (not only the 1000 words in the system vocabulary). To compare performance independently of α, we report average precision (AP), the area under the precision-recall curve as α is varied.

Table 1: *Spoken bag-of-word prediction performance (%) at two thresholds α, and the average precision (AP) over all α.*

		$\alpha = 0.4$			$\alpha = 0.7$		
Model	AP	P	R	F	P	R	F
Unigram baseline	6.8	12.1	14.2	13.1	17.6	5.9	8.8
VisionSpeechCNN	20.0	34.4	24.1	28.3	62.9	8.9	15.7
VisionSpeechPSC	18.9	40.1	20.2	26.9	62.9	6.7	12.0
OracleSpeechCNN	59.5	78.3	50.5	61.4	90.1	43.5	58.7
OracleSpeechPSC	69.7	77.1	63.0	69.3	87.4	54.1	66.8

Table 2: *Example input utterances and BoW predictions of VisionSpeechCNN for $\alpha = 0.7$. Orange shows correct predictions.*

Transcription of input utterance	Predicted BoW labels
a little girl is climbing a ladder	child, girl, little, young
a rock climber standing in a crevasse	climbing, man, rock
man on bicycle is doing tricks in an old building	bicycle, bike, man, riding, wearing
a dog running in the grass around sheep	dog, field, grass, running
a man in a miami basketball uniform looking to the right	ball, basketball, man, player, uniform, wearing
a snowboarder jumping in the air with a person riding a ski lift in the background	air, man, person, snow, snowboarder

Table 1 presents BoW prediction performance for the different models at two operating points for α, to show the trade-off between precision and recall. The unigram baseline achieves non-trivial performance, indicating that some words are commonly used across the utterances in the data set. Both VisionSpeech models substantially outperform this baseline at both α's, and in AP. Although the VisionSpeech models still lag far behind the two oracle models, the VisionSpeech models are trained without seeing any parallel speech and text. The precision of 61.3% of VisionSpeechCNN at $\alpha = 0.7$ is therefore noteworthy, since it shows that (although we miss many words in terms of recall), a relatively high-precision textual labelling system can be obtained using only images and unlabelled speech. For the oracle models, the PSC architecture is beneficial, outperforming its CNN counterpart by all measures; but for the VisionSpeech models, the PSC model falls slightly behind. We discuss this below.

Table 2 gives examples of the type of output produced by the VisionSpeechCNN model. To better analyze the model's behavior, we examine a selection of words that the model predicts that do not occur in the corresponding reference transcriptions. Figure 3 shows some of these "false alarm words", along with the most common words that do occur in the corresponding utterances. In many cases, the predicted words are variants of the correct words: e.g. for an incorrect prediction of "snow", most of the reference transcriptions contain the word "snowy". Other confusions are semantic in nature, e.g. "young" is predicted when "girl" is present, and "trick" when "ramp" is present.

[2] We do this since there are, unfortunately, some overlapping images.

[3] We also tried filterbanks; MFCCs always worked similarly or better.

[4] The code recipe is available at: `https://github.com/kamperh/recipe_vision_speech_flickr`.

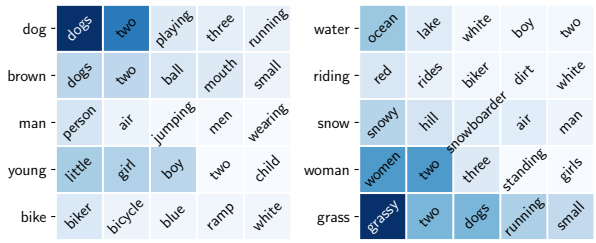

Figure 3: *False alarms (y-axes) predicted by VisionSpeechCNN, and the words in the corresponding utterances (darker shading indicating higher frequency), when used for BoW prediction.*

Table 3: *Keyword spotting performance (%).*

Model	$P@10$	$P@N$	EER
Unigram baseline	5.0	3.5	50.0
VisionSpeechCNN	54.5	33.1	22.3
VisionSpeechPSC	48.5	31.9	22.9
OracleSpeechCNN	92.0	72.4	6.2
OracleSpeechPSC	96.5	83.0	4.1

4.3. Keyword spotting

Our model can also be naturally used as a keyword spotter: given a text query, the goal is to retrieve all of the utterances in the test set containing spoken instances of that query. We randomly select 20 textual keywords from the VisionSpeech output vocabulary as queries. For evaluation, we use three metrics [47, 48]: $P@10$ is the average precision (across keywords, in %) of the 10 highest-scoring proposals; $P@N$ is the average precision of the top N proposals, with N the number of true occurrences of the keyword; and equal error rate (EER) is the average error rate at which the false acceptance and false rejection rates are equal.

Table 3 shows keyword spotting results. The trend in relative performance is similar to that of Table 1: the unigram baseline performs worst, the VisionSpeech models give reasonable scores, and the oracle models perform best. The very high oracle performance indicates that the constrained nature of the data used here (narrow domain, relatively small vocabulary) makes the task fairly easy when true transcriptions are available. Nevertheless, it is again noteworthy that both VisionSpeech models obtain a $P@10$ of around 50% at an EER of 23%, without using any text.

To give a qualitative view of VisionSpeechCNN's errors, Table 4 shows examples of incorrectly matched utterances for some keywords. As before, many of these erroneous utterances contain either variants of the keyword (e.g. "play" and "playing") or are semantically related (e.g. "young" and "little girl"). Although these matches would seem reasonable and even desirable in some settings, they are penalized under the metrics in Table 3.

4.4. Semantic keyword spotting

To investigate this issue quantitatively, we considered the top 10 proposed utterances for each keyword for each model, and relabelled as correct those utterances that either contained keyword variants or were semantically related. This allows us to report $P@10$ for the task of *semantic* keyword spotting, as shown in Table 5 (the other metrics would require us to semantically label all test utterances). Compared to Table 3, the semantic keyword spotting performance is better than exact keyword spotting scores for all models. However, the VisionSpeech models improve most, with VisionSpeechCNN improving by almost 30% absolute. Moreover, while the oracle models improved mainly due to variant matches, the VisionSpeech models had about equal numbers of relabelled variant and semantic matches.

Table 4: *Examples of incorrectly retrieved utterances when VisionSpeechCNN is used for keyword spotting.*

Keyword	Example of incorrectly matched utterance	Type
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large	. . . a rocky cliff overlooking a body of water	semantic
play	children playing in a ball pit	variant
sitting	two people are seated at a table with drinks	semantic
yellow	a tan dog jumping over a red and blue toy	mistake
young	a little girl on a kid swing	semantic

Table 5: *Semantic keyword spotting performance (%)*

Model	$P@10$
Unigram baseline	10.0
VisionSpeechCNN	82.5
VisionSpeechPSC	71.5
OracleSpeechCNN	98.0
OracleSpeechPSC	99.5

In Tables 3 and 5 we again see that while PSC is superior to CNN for the oracle models, this is not the case for the VisionSpeech models. As mentioned in §3.2, PSC is intended to also estimate word locations. Our results suggest that when trained on transcriptions (oracle models), there is a benefit in attempting to capture aspects of word order. However, when trained through visual grounding, the output of VisionSpeechPSC produces high probabilities for several semantically related words, and there is far less structure in the order of these words.

5. Conclusion

We have introduced a new way of using images to learn from untranscribed speech. By using a visual image-to-word classifier to provide soft labels for the speech, we are able to learn a neural speech-to-keyword prediction system. Our best model achieves a spoken bag-of-words precision of more than 60%, and a keyword spotting $P@10$ of more than 50% with an equal error rate of 23%. The model achieves this performance without access to any parallel speech and text. Further analysis shows that the model's mistakes are often semantic in nature, e.g. confusing "boys" and "children". To quantify this, we evaluated our model as a *semantic* keyword spotter, where the task is to find all utterances in a corpus that are semantically related to the textual keyword query. In this setting, our model achieves a semantic $P@10$ of more than 80%. Future work will consider how semantic search in speech can be formalized, and how the visual component of our approach can be explicitly tailored to obtain an improved visual grounding signal for unlabelled speech.

Acknowledgements: We thank Gabriel Synnaeve and David Harwath for assistance with data and models, as well as Shubham Toshniwal and Hao Tang for helpful feedback. This research was funded by NSF grant IIS-1433485. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

6. References

[1] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 186–197, 2008.

[2] A. Jansen *et al.*, "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proc. ICASSP*, 2013.

[3] C.-y. Lee, T. O'Donnell, and J. R. Glass, "Unsupervised lexicon discovery from acoustic input," *Trans. ACL*, vol. 3, pp. 389–403, 2015.

[4] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, "The Zero Resource Speech Challenge 2015: Proposed approaches and results," in *Proc. SLTU*, 2016.

[5] H. Kamper, A. Jansen, and S. J. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 669–679, 2016.

[6] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun.*, vol. 56, pp. 85–100, 2014.

[7] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," *arXiv preprint arXiv:1702.01991*, 2017.

[8] J. Luo *et al.*, "Object category detection using audio-visual cues," in *Proc. ICVS*, 2008.

[9] M. Sun and H. Van hamme, "Joint training of non-negative Tucker decomposition and discrete density hidden Markov models," *Comput. Speech Lang.*, vol. 27, no. 4, pp. 969–988, 2013.

[10] T. Taniguchi *et al.*, "Symbol emergence in robotics: A survey," *Adv. Robotics*, vol. 30, no. 11-12, pp. 706–728, 2016.

[11] L. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, no. 3, pp. 1558–1568, 2008.

[12] E. D. Thiessen, "Effects of visual information on adults and infants auditory statistical learning," *Cognitive Sci.*, vol. 34, no. 6, pp. 1093–1106, 2010.

[13] O. J. Räsänen, "Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions," *Speech Commun.*, vol. 54, pp. 975–997, 2012.

[14] S. Frank, N. H. Feldman, and S. J. Goldwater, "Weak semantic context helps phonetic learning in a model of infant language acquisition," in *Proc. ACL*, 2014.

[15] G. Synnaeve, M. Versteegh, and E. Dupoux, "Learning words from images and speech," in *NIPS Workshop Learn. Semantics*, 2014.

[16] D. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," in *Proc. NIPS*, 2016.

[17] T. J. Hazen, B. Sherry, and M. Adler, "Speech-based annotation and retrieval of digital photographs," in *Proc. Interspeech*, 2007.

[18] X. Anguera, J. Xu, and N. Oliver, "Multimodal photo annotation and retrieval on a mobile phone," in *Proc. ICMIR*, 2008.

[19] D. Harwath and J. R. Glass, "Learning word-like units from joint audio-visual analysis," *arXiv preprint arXiv:1701.07481*, 2017.

[20] F. Sun, D. Harwath, and J. R. Glass, "Look, listen, and decode: Multimodal speech recognition with images," in *Proc. SLT*, 2016.

[21] A. Gupta, Y. Miao, L. Neves, and F. Metze, "Visual features for context-aware speech recognition," in *Proc. ICASSP*, 2017.

[22] G. Aimetti, R. K. Moore, and L. ten Bosch, "Discovering an optimal set of minimally contrasting acoustic speech units: A point of focus for whole-word pattern matching," in *Proc. Interspeech*, 2010.

[23] V. Renkens and H. Van hamme, "Mutually exclusive grounding for weakly supervised non-negative matrix factorisation," in *Proc. Interspeech*, 2015.

[24] L. Duong *et al.*, "An attentional model for speech translation without transcription," in *Proc. NAACL*, 2016, pp. 949–959.

[25] S. Bansal, H. Kamper, A. Lopez, and S. J. Goldwater, "Towards speech-to-text translation without speech recognition," in *Proc. EACL*, 2017.

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

[27] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. CVPR*, 2015.

[28] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. CVPR*, 2015.

[29] H. Fang *et al.*, "From captions to visual concepts and back," in *Proc. CVPR*, 2015.

[30] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, 2015.

[31] C. H. Silberer, "Learning visually grounded meaning representations," Ph.D. dissertation, The University of Edinburgh, 2015.

[32] R. Bernardi *et al.*, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, pp. 409–442, 2016.

[33] A. Owens *et al.*, "Ambient sound provides supervision for visual learning," in *Proc. ECCV*, 2016.

[34] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proc. NIPS*, 2016, pp. 892–900.

[35] A. K. Vijayakumar, R. Vedantam, and D. Parikh, "Sound-Word2Vec: Learning word representations grounded in sounds," *arXiv preprint arXiv:1703.01720*, 2017.

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

[37] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. ICASSP*, 2016.

[38] K. Barnard *et al.*, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[39] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. ICCV*, 2009.

[40] M. Chen, A. X. Zheng, and K. Q. Weinberger, "Fast image tagging." in *Proc. ICML*, 2013.

[41] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. ACL*, vol. 2, pp. 67–78, 2014.

[42] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009.

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

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

[45] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: http://tensorflow.org/

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

[47] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. ASRU*, 2009.

[48] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. ASRU*, 2009.