The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Visually Grounded Cross-Lingual Keyword Spotting in Speech

Herman Kamper[1] *and Michael Roth*[2]

[1]E&E Engineering, Stellenbosch University, South Africa & [2]Saarland University, Germany

kamperh@sun.ac.za, mroth@coli.uni-sb.de

Abstract

Recent work considered how images paired with speech can be used as supervision for building speech systems when transcriptions are not available. We ask whether visual grounding can be used for *cross-lingual keyword spotting*: given a text keyword in one language, the task is to retrieve spoken utterances containing that keyword in another language. This could enable searching through speech in a low-resource language using text queries in a high-resource language. As a proof-of-concept, we use English speech with German queries: we use a German visual tagger to add keyword labels to each training image, and then train a neural network to map English speech to German keywords. Without seeing parallel speech-transcriptions or translations, the model achieves a precision at ten of 58%. We show that most erroneous retrievals contain equivalent or semantically relevant keywords; excluding these would improve P@10 to 91%.

Index Terms: visual grounding, keyword spotting, cross-lingual speech retrieval, multimodal modelling, machine translation

1. Introduction

Current automatic speech recognition (ASR) systems are trained on large amounts of transcribed speech audio. For many languages, it is difficult or impossible to collect such annotated resources [1]. Furthermore, in contrast to supervised speech systems, human infants acquire language without access to hard labels, and instead rely on other signals, such as visual cues, to ground speech [2, 3]. Recent studies have therefore started to consider how speech models can be trained on unlabelled speech paired with images [4, 5]. Grounding speech using co-occurring visual context could be a way to build systems when annotations cannot be collected, e.g. for endangered or unwritten languages [6]. In robotics, similar methods could be used to learn new words from co-occurring audio and visual signals [7].

As in [5, 6, 8], we consider the setting where unannotated images of natural scenes are paired with unlabelled spoken captions. We specifically build on [8], which proposed a model that can map speech to text labels: a trained visual tagger is used to obtain soft text labels for each training image, and a neural network is then trained to map speech to these targets. The result is a model that can be used for keyword spotting, predicting which utterances in a search collection contain a given written keyword. It does so without observing any parallel speech and text. In [8], an English visual tagger was used to ground unlabelled English speech, so English speech was searched using English keywords.

Here we propose an approach where the languages of the speech and visual tagger are not matched, with the aim of performing *cross-lingual keyword spotting*. Given a textual keyword in one language (the query language), the task is to retrieve speech utterances containing that keyword in another language (the search language). For example, given the English keyword 'doctor', the task could be to search through a spoken Swahili corpus for utterances such as *nataka kuona daktari* ('I need to see a doctor'). While parallel speech-transcriptions and translations are often difficult to obtain for low-resource languages, a collection of spoken descriptions of images could (potentially) be created by native speakers without writing or translation skills. We explore whether such paired speech-image data is sufficient for training a cross-lingual keyword spotter, thereby bringing together these two strands of research (joint image-speech modelling and cross-lingual retrieval).

Due to the lack of suitable resources in truly low-resource languages, we demonstrate a proof-of-concept implementation where we use German keywords to search through untranscribed English speech. Specifically, our setup builds on pairs of images and unlabelled English speech, and we use a visual tagger producing German text labels as targets for the speech network. For the task of cross-lingual keyword spotting, a model is given a written German keyword (e.g. *Hunde*, the German word for 'dogs') and asked to retrieve English speech utterances containing that keyword (e.g. 'two dogs playing outside near the water'). In extensive analyses, we compare the cross-lingual visual grounding model to several new alternatives (not in [8]). We find that most errors are due to semantic confusions; adjusting for these brings our model close to a directly supervised system.

2. Related work

Keyword spotting is a well-established task; the goal is to retrieve utterances in a search collection that contain spoken instances of a given written keyword [9–11]. The query and search languages are the same and typically the aim is to find exact matches. But weaker (semantic) matching has also been studied [12–15]. In cross-lingual keyword spotting, utterances in one language should be retrieved in response to user text queries in a different language. Here there has been less research, but early work [16] proposed to cascade ASR with text-based cross-lingual information retrieval [17]. This is only possible when transcribed speech are available for building an ASR system. Some recent work has proposed models that can translate speech in one language directly to text in another [18–21], but this requires parallel speech with translated text. We use visual context as supervision for settings where translations are unavailable.

Several recent studies have trained models on images paired with unlabelled speech [4–6, 22–26]. Most approaches map images and speech into a common space, allowing images to be retrieved using speech and vice versa. Although useful, labelled (textual) predictions are not possible. The model of [8] uses an external visual tagger to tag training images with text labels, enabling the model to map speech to text labels (without using transcriptions). We extend this approach by applying a visual tagger in one language to parallel images and speech in another (the search) language. To our knowledge, cross-lingual keyword spotting has not been attempted using visual speech grounding. Finally, recent work has used vision as an additional input modality for (textual) machine translation [27–29]. We consider *speech* retrieval, with vision as the *only* supervisory signal.

10.21437/SLTU.2018-52

3. Model

Given an unlabelled corpus of parallel images and spoken captions in the search language (English), we use an external visual tagger in the query language (German) to produce soft targets for a speech network. This is illustrated in Figure 1, where training image I is paired with English caption $X = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$, with each frame $\boldsymbol{x}_t$ an acoustic feature vector, e.g. Mel-frequency cepstral coefficients (MFCCs). Image I is tagged with German text labels, which serves as targets for the speech network $\boldsymbol{f}(X)$. The result is a network that maps English speech to German keyword labels (ignoring order, quantity and where the translations of the keywords occur). During testing, the model is applied as a cross-lingual keyword spotter as shown in Figure 2; each speech utterance in an unseen English search collection is passed through $\boldsymbol{f}(X)$, and the output is used to predict whether a given German keyword (text query) is present. In testing, only English speech input is used (no images). We now give full details.

3.1. Detailed model description

For training (Figure 1), if we knew the German words occurring in English training utterance X, we could construct a multi-hot cross-lingual bag-of-word (BoW) vector $\boldsymbol{y}_{\text{xbow}} \in \{0, 1\}^W$, with W the vocabulary size and each dimension $y_{\text{xbow},w}$ a binary indicator for whether X contains a translation of German word w. However, we do not have transcriptions or translations to obtain such ideal cross-lingual BoW supervision. Instead, we only have the image I which is paired with X. Rather than binary indicators, we use a multi-label visual tagging system producing soft targets $\hat{\boldsymbol{y}}_{\text{de}} \in [0, 1]^W$, with $\hat{y}_{\text{de},w} = P(w|I)$ the estimated probability of German word w being relevant given image I. In Figure 1, $\hat{\boldsymbol{y}}_{\text{de}}$ would ideally be close to 1 for w corresponding to words such as *Feld* (field), *Hunde* (dogs), *springt* (jump), and *grün* (green), and close to 0 for irrelevant dimensions. Note that the visual tagger is assumed to be external: whereas the speech network $\boldsymbol{f}(X)$ is trained, the tagger is held constant.

Given $\hat{\boldsymbol{y}}_{\text{de}}$ as target, we train the speech model $\boldsymbol{f}(X)$ (Figure 1, right). This model (parameters $\boldsymbol{\theta}$) consists of a convolutional neural network (CNN) over the speech X with a final sigmoidal layer so that $\boldsymbol{f}(X) \in [0, 1]^W$. We interpret each dimension of the output as $f_w(X) = P_{\boldsymbol{\theta}}(w|X)$. We train $\boldsymbol{f}(X)$ using the summed cross-entropy loss, which (for a single training example) is:

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

If we had $\hat{y}_{\text{de},w} \in \{0, 1\}$, as in $\boldsymbol{y}_{\text{xbow}}$, this would be the summed log loss of W binary classifiers. Note that the size-W (German) vocabulary of the system is implicitly given by the visual tagger.

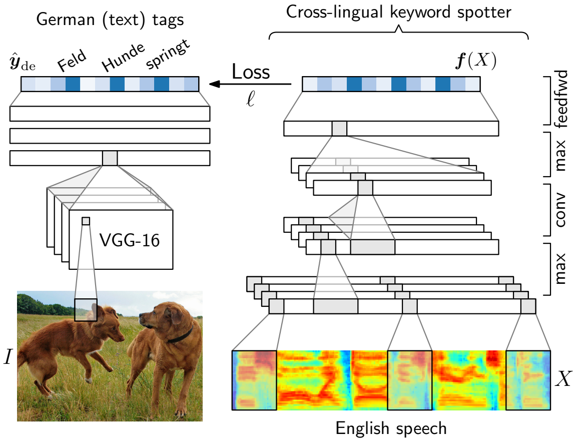

Figure 1: *During training, an external German visual tagger produces text targets for a speech network taking in English speech. The speech network is therefore trained using only parallel images and unlabelled English spoken captions. Here the visual target $\hat{\boldsymbol{y}}_{\text{de}}$ should ideally be close to 1 for words such as Feld* (field)*, Hunde* (dogs)*, springt* (jump)*, and grün* (green)*.*

3.2. The German visual tagger

A visual tagger is a multi-label computer vision system that predicts an unordered set of words (nouns, adjectives, verbs) that accurately describes aspects of a scene [30–32]. Ideally we want an existing vision system in the query language (Figure 1, left). Although it is fair to assume such a system would be available if the query language is high resource (§1), we could not find an off-the-shelf German tagger. We therefore train our own German visual tagger on separate data.

We use the Multi30k dataset, which contains around 30k images each annotated with five written German captions [33]. Captions are combined into a single BoW target after removing stop words. As basis for our tagger, we use VGG-16 [34], trained on around 1.3M images [35], but we replace the final classification layer with four 2048-unit ReLU layers followed by a final sigmoidal layer for predicting word occurrence. VGG-16 was used in a similar way in previous vision-speech models [5,6]. The visual tagger is trained on Multi30k with the output layer limited to the $W = 1\text{k}$ most common German word types in the captions. Only the additional fully-connected layers are updated, i.e. the VGG-16 parameters are not fine-tuned.

Importantly, none of the training images here overlap with the test data used in our experiments (§4). Thus, the visually grounded model does not get even indirect access to the (written) German translations, so we use it as if it is external.

4. Experiments

Our goal is to find spoken utterances in a search language that contain written keywords from a query language. Our model, referred to as XVisionSpeechCNN, does so without using any transcribed or parallel data; instead, it relies solely on utterances

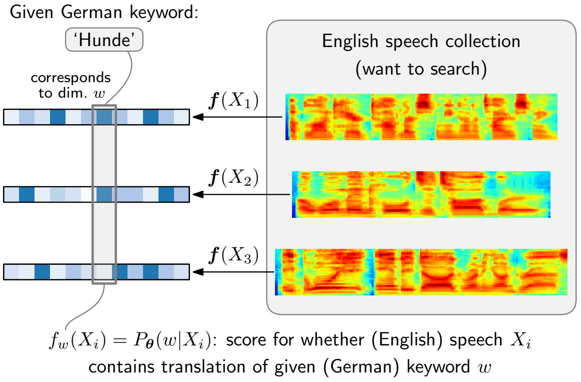
Given German keyword:
'Hunde'
corresponds to dim. w
English speech collection (want to search)
$\boldsymbol{f}(X_1)$
$\boldsymbol{f}(X_2)$
$\boldsymbol{f}(X_3)$
$f_w(X_i) = P_{\boldsymbol{\theta}}(w|X_i)$: score for whether (English) speech X_i contains translation of given (German) keyword w

Figure 2: *After training, $\boldsymbol{f}(X)$ can be applied as a cross-lingual keyword spotter. An unseen English test utterance is passed through the model, and the resulting output is interpreted as a score for a particular German keyword occurring in the speech.*

in the search language that are paired with images, which we process by automatically adding visual keywords in the query language (§3, Figure 1). At test time, our proposed model is given a keyword in the query language and has to retrieve corresponding utterances in the search language, without access to parallel data, transcriptions, or utterance-image pairs (Figure 2).

4.1. Experimental setup and evaluation

We train our visually grounded cross-lingual keyword spotting model XVisionSpeechCNN on the Flickr8k Audio Captions Corpus of parallel images and spoken captions containing 8k images, each with five spoken English captions [36]. The audio comprises around 37 hours of non-silent speech, and comes with train, development and test splits of 30k, 5k and 5k utterances, respectively. We parameterise speech as 13 MFCCs with first and second order derivatives, giving 39-dimensional input vectors. Utterances longer than 8 s are truncated (99.5% are shorter). XVisionSpeechCNN has the same structure as the monolingual model from [25]: 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; 3k-unit fully-connected ReLU; and the 1k-unit sigmoid output. We train using Adam [37] for 25 epochs with early stopping, a learning rate of $1 \cdot 10^{-4}$ and a batch size of eight.

For evaluation, we need reference German translations for the English test utterances. The data set of [33] contains German translations for a subset of the English development and test utterance of the Flickr8k corpus. We perform evaluation on these utterances, with approximately 1k English utterances for development and 1k utterances for testing, each with one German reference translation. As keywords, we randomly selected 39 words from the data on which the visual tagger was trained.

When applying XVisionSpeechCNN to test data (Figure 2), we interpret the output $f_w(X) \in [0, 1]$ as a score for how relevant an English utterance X is given German keyword w. The models we compare to (below) also gives this type of scoring. To obtain a hard prediction from a model, we set a threshold α and label all keywords for which $f_w(X) > \alpha$ as relevant. By comparing this to the reference translation, precision and recall can be calculated. We stem the words in both the prediction and reference translation, so that inflections are not marked as errors. To measure performance independent of α, we report *average precision (AP)*, the area under the precision-recall curve as α is varied. We also consider how a model ranks utterances in the test data from most to least relevant for each keyword [38, 39]: *precision at ten (P@10)* is the average precision of the ten highest-scoring proposals; *precision at N (P@N)* is the average precision of the top N proposals, with N the number of true occurrences of translations of the word; and *equal error rate (EER)* is the rate at which false acceptance and rejection rates are equal.

4.2. Baselines and comparison models

DETextPrior. This baseline completely ignores the search language utterance and relies only on unigram probabilities of the keywords in the query language. Comparisons to DETextPrior indicate how much better our model does than simply predicting common German words for any English speech input.

DEVisionCNN. One question is whether XVisionSpeechCNN learns to ignore aspects of the acoustics that are not indicative of visual targets. This baseline is an attempt to test this: as the representation for each test utterance, it passes through the German visual tagger the *true* image paired with that utterance. If XVisionSpeechCNN had access to ideal visual tags, then DEVisionCNN would be an upper bound, but in reality our model could do better or worse (since training does not generalise perfectly).

XBoWCNN. To check how reliable automatically predicted tags are in comparison to ground truth text labels, we train as an upper bound XBoWCNN, which has access to the keywords that indeed appear as translations in the search language utterances.

4.3. Results

To first illustrate the cross-lingual keyword spotting task, Table 1 shows example output from XVisionSpeechCNN for a selection of German keywords with the top English utterances that were retrieved in each case. Utterances where the reference German translation did not contain the given keyword are marked with $*$. Of the 24 shown retrievals, ten are incorrect.

Table 1: *Transcriptions of the top English utterances retrieved using* XVisionSpeechCNN *for a selection of German keywords. Incorrect retrievals (i.e. where the reference translation of the utterance does not contain the keyword) are marked with $*$. Different error types (Table 3) are marked with (1), (2) and (3).*

Fahrrad (bicycle)	
man riding a bicycle on a foggy day	
a biker does a trick on a ramp $*$	(2)
a person is doing tricks on a bicycle in a city	
Feld (field)	
a team of baseball players in blue uniforms walking together on field	
a brown and black dog running through a grassy field $*$	(1)
two small children walk away in a field	
groß(en) (big)	
a large crowd of people ice skating outdoors	
a surfer catching a large wave in the ocean	
a small group of people sitting together outside $*$	(3)
grün(en) (green)	
boy wearing a green and white soccer uniform running through the grass	
a girl is screaming as she comes off the water slide $*$	(3)
a brown dog is chasing a red frisbee across a grassy field $*$	(2)
Hemd (shirt)	
a woman in a red shirt and a man in white stand in front of a mirror	
a man in a blue shirt lifts up his tennis racket and smiles	
a man in blue cap and jacket looks frustrated $*$	(2)
klettern/klettert (climbing)	
a lone rock climber in a harness climbing a huge rock wall	
a man is rock climbing at sunset $*$	(1)
a man is laying under a large rock in the forest $*$	(2)
Personen (people)	
two people are riding a ski lift with mountains behind them	
two women are climbing over rocks near to the ocean $*$	(2)
two people sit on a bench leaned against a building with writing on it	
Straße (street)	
a woman in black and red listens to an ipod walks down the street	
people on the city street walk past a puppet theater	
an asian woman rides a bicycle in front of two cars $*$	(2)

Table 2: *Cross-lingual keyword spotting results (%) on test data.*

Model	$P@10$	$P@N$	EER	AP
DETextPrior	7.2	6.3	50	10.4
DEVisionCNN	41.5	32.9	25.9	29.7
XVisionSpeechCNN	58.2	40.4	23.5	40.0
XBoWCNN	80.8	54.3	19.1	54.3

Table 3: *Analysis of errors by a human annotator of the top ten retrievals on development data. Percentages (%) indicate the absolute drop in $P@10$ due to that error type.*

	XVisionSpeech		XBoWCNN	
Error type	Count	%	Count	%
(1) Correct (exact)	32	8.2	45	11.5
(2) Semantically related	86	22.1	13	3.3
(3) Incorrect retrieval	35	9.0	19	4.9
Total	153	39.3	77	19.7

Table 2 shows the results on the test data for XVisionSpeechCNN and the upper- and lower-bound models. Without seeing any speech transcriptions or translated text, XVisionSpeechCNN achieves a $P@10$ of 58%, with XBoWCNN the only model to outperform the visually grounded model. By comparing performance to DETextPrior, we see that XVisionSpeechCNN is not just predicting common German words. Interestingly, XVisionSpeechCNN also outperforms DEVisionCNN over all metrics. If the former were perfectly predicting the German visual tags (which is what it is trained to do), then the performance of these two models would be the same. We see, however, that XVisionSpeechCNN is doing more than simply mapping the acoustics to the visual tags; we speculate that it is therefore picking up information in the speech which cannot be obtained from the corresponding test images.

4.4. Further analysis

Error analysis. Around 40% of the utterances in the top ten retrievals of XVisionSpeechCNN still do not contain the given German keyword in the reference translation. To understand the nature of these mistakes, we asked a German native speaker to annotate each error in the top ten retrievals with one of the following categories: (1) the reference does not contain the keyword literally, but an equivalent translation; (2) the utterance does not contain a translation of the keyword, but the retrieval is related in meaning; or (3) the retrieval is completely incorrect. Examples of the three types of errors are marked on the right in Table 1. Errors of type (1) are normally due to a synonym being used; e.g. in Table 1 the erroneous utterance shown for *Feld* (field) is a plausible retrieval as the reference contains the word *Wiese* (meadow). An example error of type (2) can be seen for the keyword *Hemd* (shirt): here, the retrieved utterance does not contain the keyword, but mentions other clothing (cap, jacket).

Errors from both XVisionSpeechCNN and XBoWCNN were presented to the annotator in shuffled order. Table 3 indicates the absolute penalty in $P@10$ for each error type on development data. For both models, around 10% of the retrievals marked as errors are actually correct. The bulk of errors from XVisionSpeechCNN is due to semantically related retrievals. These retrievals are marked as errors, but could actually be useful depending on the type of retrieval application. This is in line with [25], which showed that visual supervision is beneficial for retrieving non-exact but still relevant utterances in the monolingual case. If type (1) and type (2) errors are not counted as incorrect, XVisionSpeechCNN and XBoWCNN would achieve a $P@10$ of 91% and 95%, respectively (but, again, this will depend on the use-case). We leave a larger analysis, which will also measure recall (not only top retrievals), for future work.

Table 4: *Cross-lingual keyword spotting results (%) for different variants of* XVisionSpeechCNN *on development data.*

Model	$P@10$	$P@N$	EER	AP
XVisionSpeechCNN	60.8	39.3	23.1	38.0
KeyXVisionSpeechCNN	60.0	39.6	24.5	36.9
OracleXVisionSpeechCNN	57.4	37.6	24.8	36.5

Variants and ideal supervision. We compare different variants of XVisionSpeechCNN to gain insight into properties of the model. XVisionSpeechCNN produces scores $\boldsymbol{f}(X) \in [0,1]^W$ for all $W = 1\text{k}$ words in its output vocabulary. But we are actually only interested in those dimensions w corresponding to the test keywords. If we knew the keywords at training time, we could train a model which only tries to predict the visual tags corresponding to these keywords. Table 4 shows development performance for such a model, KeyXVisionSpeechCNN. Performance is similar to that of XVisionSpeechCNN, with the latter being slightly better on most metrics. To understand this improvement, note that XVisionSpeechCNN can be seen as a variant of KeyXVisionSpeechCNN trained in a multitask fashion: it is trying to predict extra words not used during testing [40]. This effectively regularises our model (improving results).

XVisionSpeechCNN is trained on soft scores from a visual tagger. What if we had the true hard assignments from the manual annotations for the training images? OracleXVisionSpeechCNN is trained on such oracle targets. Table 4 shows that this is actually detrimental. In [41], where video was paired with general audio (not speech), soft targets were also used (as in XVisionSpeechCNN). They described this as a student-teacher approach, where the student (in our case the speech network) is trying to distil knowledge from the teacher network (in our case the visual tagger). It has been shown [42, 43] that training using soft targets can be beneficial for the student network, which aligns with our findings here.

5. Conclusion

We proposed the first visually grounded speech model for cross-lingual keyword spotting. By labelling images with tags from a multi-label vision system in the query language (German), we train a network that maps unlabelled speech in the search language (English) to German keyword labels. Using this network for spotting whether translations of German keywords occur in English speech, we achieve a $P@10$ of almost 60%. The majority of errors are due to semantically related retrievals; when these are taken into account, our approach comes close to a supervised model trained on parallel speech with text translations. In further analysis, we showed that by implicitly predicting tags not in the keyword set, we are getting a small benefit from multitask learning. We also showed that using soft targets from the visual tagger is better than oracle hard targets; this aligns with findings in student-teacher knowledge distillation studies. Future work will consider error analyses at a larger scale and applications on truly low-resource (e.g. unwritten) languages.

6. References

[1] 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.

[2] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Sci.*, vol. 26, no. 1, pp. 113–146, 2002.

[3] O. Räsänen and H. Rasilo, "A joint model of word segmentation and meaning acquisition through cross-situational learning," *Psychol. Rev.*, vol. 122, no. 4, pp. 792–829, 2015.

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

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

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

[7] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh, "Symbol emergence in robotics: A survey," *Adv. Robotics*, vol. 30, no. 11-12, pp. 706–728, 2016.

[8] H. Kamper, S. Settle, G. Shakhnarovich, and K. Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," *Proc. Interspeech*, 2017.

[9] J. G. Wilpon, L. R. Rabiner, C.-H. Lee, and E. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden markov models," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, no. 11, pp. 1870–1878, 1990.

[10] I. Szöke, P. Schwarz, P. Matejka, L. Burget, M. Karafiát, M. Fapso, and J. Cernockỳ, "Comparison of keyword spotting approaches for informal continuous speech," in *Proc. Interspeech*, 2005.

[11] A. Garcia and H. Gish, "Keyword spotting of arbitrary words using minimal speech resources," in *Proc. ICASSP*, vol. 1. IEEE, 2006, pp. I–I.

[12] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Proc. Mag.*, vol. 25, no. 3, 2008.

[13] H.-Y. Lee, T.-H. Wen, and L.-S. Lee, "Improved semantic retrieval of spoken content by language models enhanced with acoustic similarity graph," in *Proc. SLT*, 2012.

[14] Y.-C. Li, H.-y. Lee, C.-T. Chung, C.-a. Chan, and L.-s. Lee, "Towards unsupervised semantic retrieval of spoken content with query expansion based on automatically discovered acoustic patterns," in *Proc. ASRU*, 2013.

[15] L.-s. Lee, J. Glass, H.-y. Lee, and C.-a. Chan, "Spoken content retrieval—beyond cascading speech recognition with text retrieval," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 1389–1420, 2015.

[16] P. Sheridan, M. Wechsler, and P. Schäuble, "Cross-language speech retrieval: Establishing a baseline performance," in *Proc. SIGIR*, 1997.

[17] D. W. Oard and A. R. Diekema, "Cross-language information retrieval," *ARIST*, vol. 33, pp. 223–56, 1998.

[18] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proc. NAACL*, 2016, pp. 949–959.

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

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

[21] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," *Proc. ICASSP*, 2018.

[22] J. Drexler and J. Glass, "Analysis of audio-visual features for unsupervised speech recognition," 2017.

[23] K. Leidal, D. Harwath, and J. Glass, "Learning modality-invariant representations for speech and images," *Proc. ASRU*, 2017.

[24] O. Scharenborg *et al.*, "Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "Speaking Rosetta" JSALT 2017 workshop," *arXiv preprint arXiv:1802.05092*, 2018.

[25] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic keyword spotting by learning from images and speech," *arXiv preprint arXiv:1710.01949*, 2017.

[26] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," *Proc. ICASSP*, 2018.

[27] L. Specia, S. Frank, K. Sima'an, and D. Elliott, "A shared task on multimodal machine translation and crosslingual image description," in *Proc. WMT*, 2016.

[28] D. Elliott and A. Kádár, "Imagination improves multimodal translation," *arXiv preprint arXiv:1705.04350*, 2017.

[29] D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia, "Findings of the second shared task on multimodal machine translation and multilingual image description," in *Proc. WMT*, 2017.

[30] K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[31] 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.

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

[33] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual English-German image descriptions," in *Proc. Workshop Vision Language*, 2016.

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

[35] 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.

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

[37] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

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

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

[40] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

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

[42] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. CVPR*, 2016.

[43] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.