Semantic speech retrieval with a visually grounded model of untranscribed speech

Herman Kamper[1], Gregory Shakhnarovich[2], Karen Livescu[2]
[1]E&E Engineering, Stellenbosch University & [2]Toyota Technological Institute at Chicago
kamperh@sun.ac.za, greg@ttic.edu, klivescu@ttic.edu

Abstract

There is growing interest in speech models that can learn from unlabelled speech paired with visual context. Here we study how a visually grounded speech model, trained on images of scenes paired with spoken captions, captures aspects of semantics. We use an external image tagger to generate soft text labels from images, which serve as targets for a neural model that maps untranscribed speech to (semantic) keyword labels. We introduce a newly collected data set of human semantic relevance judgements and an associated task, semantic speech retrieval, where the goal is to search for spoken utterances that are semantically relevant to a given text query. Without seeing any text, the model trained on parallel speech and images achieves a precision of almost 60% on its top ten semantic retrievals. Compared to a supervised model trained on transcriptions, our model matches human judgements better by some measures, especially in retrieving non-verbatim semantic matches.

1. Introduction

Current methods for automatic speech recognition (ASR) require large amounts of transcribed speech data. This has prompted work on models that, instead of using exact transcriptions, can learn from weaker or noisy forms of supervision. Our work here builds on a line of recent studies [4, 7, 10, 16, 17] that use natural images of scenes paired with untranscribed spoken descriptions. Neither the spoken nor visual input is labelled. This setting is relevant for low-resource speech processing [4], robotics [19], and human language acquisition research [14, 15].

Most approaches map images and speech into some common space, allowing images to be retrieved using speech and vice versa. Although useful, such models cannot predict (written) labels for the input speech. Here we specifically analyse a model based on Kamper et al. [10], which can make text label predictions. A trained visual tagger is used to obtain soft text labels for each training image, and these are used as targets for a neural network that maps speech to keyword labels. Without observing any transcriptions, the model of [10] was used as a keyword spotter, predicting which utterances in a search collection contain a given written keyword. It was observed that the model often confuses semantically related words; these count as errors in keyword spotting, but could be useful in *semantic* search applications.

Our primary aim here is to perform an extensive analysis to see what aspects of semantics are captured by the model of [10]. To do so formally, we use the task of *semantic speech retrieval*, where the aim is to retrieve all utterances in a speech collection that are semantically relevant to a given query keyword, irrespective of whether that keyword occurs exactly in an utterance or not. E.g., given the query 'children', the goal is to return not only utterances containing the word 'children', but also utterances about children, like 'young boys playing soccer in the park'. There has been some work on this and related tasks (see [11] for a complete review), but typically in higher-resource settings and none using visual supervision. Using a newly collected corpus of speech data with soft semantic relevance judgements, we present an extensive analysis of an updated version of the model of [10], and compare it to several new alternative models for the task of semantic speech retrieval.

2. The visually grounded speech model

Given a corpus of parallel images and spoken captions, neither with textual labels, we train a spoken keyword prediction model using a visual tagging system to produce soft labels for the speech network. The overall approach is illustrated in Figure 1. Training image I is paired with a spoken caption $X = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$, where each frame $\boldsymbol{x}_t$ is an acoustic feature vector, e.g. Mel-frequency cepstral coefficients (MFCCs). We use an external vision system to tag I with soft textual labels, giving $\hat{\boldsymbol{y}}_{\text{vis}} \in [0, 1]^W$ where $\hat{y}_{\text{vis},w} = P_{\boldsymbol{\gamma}}(w|I)$ is the estimated probability of word w being present in image I under vision model parameters $\boldsymbol{\gamma}$. These tagger outputs are then used as targets to train

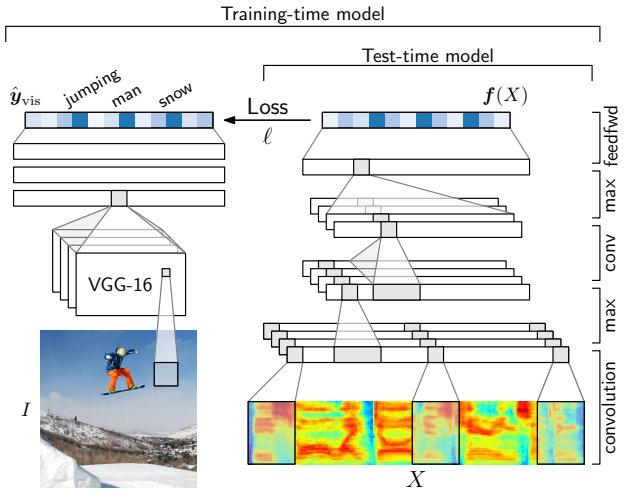

Figure 1. For training, an external visual tagger produces soft tags for image I, which serve as targets for the speech network fed with spoken caption X. In testing, the speech network is given unseen speech (no image) and the output $f(X)$ is used for semantic retrieval of a textual keyword.

the speech network $\boldsymbol{f}(X)$ (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 $\boldsymbol{f}(X) \in [0,1]^W$, where $f_w(X) = P_{\boldsymbol{\theta}}(w|X)$ is interpreted as the posterior probability of tag w given the spoken utterance. We train the speech model using the summed cross-entropy loss:

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

The resulting network $\boldsymbol{f}(X)$ can then predict which keywords are present for a given utterance X, disregarding the order, quantity, or locations of the keywords in the input speech. The possible keywords (the vocabulary) are implicitly specified by the visual tagger. No transcriptions are used during training. When applying the trained $\boldsymbol{f}(X)$ for keyword spotting or semantic retrieval, test speech alone is used without any visual input.

Our approach requires a visual tagging system [2, 3, 5] that predicts an unordered set of words that describe the scene (Figure 1, left). We train our visual tagger on combined data from the Flickr30k [22] and MSCOCO [12] data sets, which consist of images each with five written captions. For each image, the training labels consist of a bag-of-words vector $\boldsymbol{y}_{\text{vis}} \in [0,1]^W$ containing indicators for the $W = 1000$ most common content words in the image captions. None of the images used here occur in the parallel image-speech data used in our experiments.

3. Semantic speech retrieval data set

We consider the task of *semantic speech retrieval*. Instead of matching keywords exactly, as is typical in keyword spotting [18, 20], the aim is to retrieve all utterances that are semantically relevant, irrespective of whether the keyword occurs in the utterance or not. E.g., for the query 'sidewalk', a model should return not only utterances containing the word exactly, but also speech like 'an old couple window-shopping on a Paris street.'

We collect a new data set of human semantic judgements by extending the corpus of [6], which consists of parallel images and spoken captions. The data is transcribed, but not semantically labelled. For a subset of the speech in the corpus, we use Amazon Mechanical Turk (AMT) to collect semantic labels from human annotators. The annotators choose which of a set of keywords could be used to search for a given utterance describing a scene. Five workers annotate each utterance.

To evaluate a semantic keyword retrieval model against the human annotations, one option is to combine the human judgements into a single hard label based on the majority decision. On the other hand, we found there was a wide range of opinions among the human annotators, indicating that semantic relevance may be inherently "soft". This motivates evaluation by comparing against the proportion of annotators that agree with a given label. In our experiments we consider both hard and soft options.

4. Experimental setup and evaluation

We train our model on the corpus of parallel images and spoken captions of [6], containing 8000 images with five spoken captions each. Audio comprises around 37 hours of active speech. We parameterise the speech audio as 13 MFCCs with first and second order derivatives, giving 39-dimensional input vectors. Training images are passed through the visual tagger, producing soft targets $\hat{\boldsymbol{y}}_{\text{vis}}$ for training the keyword prediction model $\boldsymbol{f}(X)$ on the unlabelled speech, as shown in Figure 1. We refer to the resulting model as VISIONSPEECHCNN. The structure and training procedure used for VISIONSPEECHCNN is very similar to that of [10] (see that paper for architectural details).

4.1. Evaluation

To use VISIONSPEECHCNN for semantic retrieval, we use its output $f_w(X) \in [0,1]$ as a score for how relevant an utterance X is given the keyword w. The baseline and fully supervised models below similarly predict a relevance score for each utterance given a specific keyword.

We compare a model's predictions to semantic labels obtained from human annotators using several metrics. To obtain a hard labelling from a model, we set a threshold α, and label all keywords for which $f_w(X) > \alpha$ as relevant. By

Table 1. Keyword spotting and semantic speech retrieval performance for VISIONSPEECHCNN (row 3), compared against the baseline (rows 1 and 2) and fully supervised (rows 4 and 5) models. Boldface indicates both the top-scoring models that does not use transcriptions (rows 1 to 3) as well as the best supervised model (rows 4 and 5) for each of the metrics.

	Exact keyword spotting (%)				Semantic speech retrieval (%)				
Model	P@10	P@N	EER	AP	P@10	P@N	EER	AP	Spear. ρ
Baseline models:									
1. TEXTPRIOR	2.8	3.4	50.0	8.7	6.1	7.0	50.0	11.4	10.8
2. VISIONTAGPRIOR	2.8	3.4	50.0	7.0	6.1	7.0	50.0	13.6	12.5
3. VISIONSPEECHCNN	**38.5**	**30.8**	**19.6**	**26.9**	**58.8**	**39.7**	**23.9**	**39.4**	**32.4**
Fully supervised models:									
4. SUPERVISEDBOWCNN	**84.9**	**74.7**	5.6	**87.3**	**88.1**	50.3	23.8	51.3	21.9
5. TEXTWUP	65.4	67.3	**2.6**	75.2	80.3	**63.0**	**19.4**	**60.9**	25.2

comparing this to the ground truth semantic labels (according to majority annotator agreement), precision and recall can be calculated; we report **average precision (AP)**, the area under the precision-recall curve as α is varied. The soft scores $f_w(X)$ can also be compared directly to the number of annotators that selected the keyword w for utterance X: we use **Spearman's** ρ to measure the correlation between the rankings of these two variables, as is common in work on word similarity [1, 9]. The remaining metrics are standard in (exact) keyword spotting, based on how a model ranks utterances in the test data from most to least relevant for each keyword [8, 23]: **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 the keyword; and **equal error rate (EER)** is the average error rate at which false acceptance and rejection rates are equal.

4.2. Baselines and fully supervised models

TEXTPRIOR uses the unigram probability of each keyword estimated from the transcriptions of the training portion of the spoken captions corpus. This will indicate how much better our model does than simply hypothesising common words. Similarly, VISIONTAGPRIOR is obtained by passing all training images through the trained visual tagger and then taking the average over all images.

The SUPERVISEDBOWCNN fully supervised model uses transcriptions to obtain hard bag-of-words (BoW) supervision: $\boldsymbol{y}_{\text{bow}}$ targets are constructed for the 1000 most common content words in the transcriptions of the training utterances (ignoring stop words).

Suppose we had a perfect ASR system, converting input speech to text without errors. How well could we do at semantic speech retrieval using this error-free text? To answer how this cascaded approach would do in an ideal setting, we consider a text-based semantic retrieval method using transcriptions of the speech. WuP similarity, named after Wu and Palmer [21], scores the semantic relatedness between two words according to the path length between them in the WordNet lexical hierarchy [13]. Our TEXTWUP upper-bound model is based on the closest WuP match between a keyword and each of the words in a transcribed utterance.

5. Experimental results and analysis

Exact and semantic keyword retrieval performance for VISIONSPEECHCNN and all the baseline and fully supervised models are shown in Table 1. VISIONSPEECHCNN outperforms the baseline models across all metrics, indicating that it does more than simply outputting common labels.

The baseline models and VISIONSPEECHCNN all perform better at semantic than at exact retrieval. In contrast, the transcription-based supervised models (rows 4 and 5) perform better on P@10, but worse on all other semantic search metrics. P@10 only measures precision of the highest ranked utterances, while the other metrics combine precision and recall; thus, the transcription-based supervised models struggle to retrieve semantic matches compared to exact matches, while VISIONSPEECHCNN recall more semantic matches. In terms of absolute performance, the transcription-based models still perform better at semantic speech retrieval on the metrics based on hard ground truth labels. However, for Spearman's ρ, which gives credit even if a prediction does not match the majority of annotations, VISIONSPEECHCNN outperforms all other models. Visual context is clearly beneficial in matching soft human ratings.

Despite the benefit of visual supervision, SUPERVISEDBOWCNN still performs better on semantic speech retrieval measured against hard labels. We analysed these results and found that VISIONSPEECHCNN actually performs better on non-verbatim semantic matches. On the metrics that use hard ground-truth labels, TEXTWUP performs best. However, VISIONSPEECHCNN again performs better on Spearman's ρ, which measures the ranking correlation with annotator counts. As noted, the visually trained models are par-

ticularly strong in matching non-exact semantic keywords.

6. Conclusion

We investigated how a model that learns from parallel images and unlabelled speech captures aspects of semantics in speech. We collected a new data set for a semantic speech retrieval task, where the aim is to retrieve utterances that are semantically relevant given a written query keyword. Without seeing any parallel speech and text, the visually grounded model achieves a semantic $P@10$ of almost 60%. Although a model trained on transcriptions is superior on some metrics, the vision-speech model retrieves more non-verbatim semantic matches and is a better predictor of the actual soft human ratings.

References

[1] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proc. HLT-NAACL*, 2009.

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

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

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

[5] 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.

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

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

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

[9] F. Hill, R. Reichart, and A. Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Comput. Linguist.*, 41(4), 2015.

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

[11] L.-s. Lee, J. R. Glass, H.-y. Lee, and C.-a. Chan. Spoken content retrieval—beyond cascading speech recognition with text retrieval. *IEEE Trans. Audio, Speech, Language Process.*, 23(9):1389–1420, 2015.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proc. ECCV*, 2014.

[13] G. A. Miller. WordNet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[14] O. Räsänen and H. Rasilo. A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychol. Rev.*, 122(4):792–829, 2015.

[15] D. K. Roy and A. P. Pentland. Learning words from sights and sounds: A computational model. *Cognitive Sci.*, 26(1):113–146, 2002.

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

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

[18] 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.

[19] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh. Symbol emergence in robotics: A survey. *Adv. Robotics*, 30(11-12):706–728, 2016.

[20] 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.*, 38(11):1870–1878, 1990.

[21] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proc. ACL*, 1994.

[22] 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*, 2:67–78, 2014.

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