*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Towards Bilingual Lexicon Discovery From Visually Grounded Speech Audio

*Emmanuel Azuh, David Harwath, James Glass*

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
{emazuh, dharwath, glass}@mit.edu

## Abstract

In this paper, we present a method for the discovery of word-like units and their approximate translations from visually grounded speech across multiple languages. We first train a neural network model to map images and their spoken audio captions in both English and Hindi to a shared, multimodal embedding space. Next, we use this model to segment and cluster regions of the spoken captions which approximately correspond to words. Finally, we exploit between-cluster similarities in the embedding space to associate English pseudo-word clusters with Hindi pseudo-word clusters, and show that many of these cluster pairings capture semantic translations between English and Hindi words. We present quantitative cross-lingual clustering results, as well as qualitative results in the form of a bilingual picture dictionary.

**Index Terms**: Low-resource speech processing, multimodal speech processing, cross-lingual speech processing

## 1. Introduction

With the many languages in the world, people often need to cross language barriers to communicate. Researchers have made huge strides towards making automated machine translation more and more reliable. Current speech-to-speech translation systems rely on a cascade of models that perform automatic speech recognition, machine translation, and text-to-speech synthesis [1]. These models each require large quantities of manually-annotated training data, but transcribing parallel corpora of speech audio in both the source and target languages can be prohibitively costly. The text bottleneck also makes it difficult to automatically translate to and from languages without a written orthography. In this work, we attempt to align semantically equivalent words across languages directly at the speech signal level, without the need for text transcripts. We build on the work presented by [2, 3, 4, 5], which showed that multimodal neural network models could be trained to directly associate speech waveforms with images, resulting in the ability to recognize spoken words in continuous speech signals without the need for conventional ASR. This type of model was generalized to handle speech inputs from two different languages in [5], and was shown to be capable of cross-lingual matching of semantically similar captions. We take the work in [5] a step further by explicitly locating and clustering the words learned by the model in both languages. We then semantically link the discovered English clusters with the Hindi clusters (Figure 1), as well as image regions that they are most similar to. This forms the basis of a picture dictionary, which shows segments of speech from both languages coupled with semantically relevant regions of images.

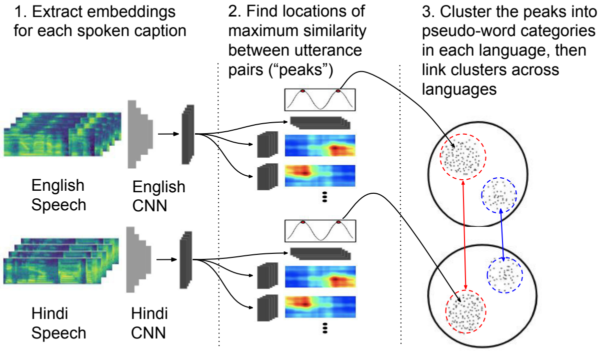

1. Extract embeddings for each spoken caption
2. Find locations of maximum similarity between utterance pairs ("peaks")
3. Cluster the peaks into pseudo-word categories in each language, then link clusters across languages
English Speech
English CNN
Hindi Speech
Hindi CNN

Figure 1: *Speech CNN embeddings (left) are compared to each other using dot product to find and extract regions of high similarity from utterances containing similar concepts (middle). The embeddings at these regions are clustered separately for each language and linked using cross-lingual cluster centroid similarity (right).*

## 2. Prior work

The high cost of current supervised methods of speech recognition and machine translation has led to several proposed methods for increasingly unsupervised speech recognition. Segmental Dynamic Time Warping proposed by [6] and extended in [7, 8, 9] operates on raw audio to find patterns within speech utterances to automatically discover word categories. Other works such as [10, 11, 12] used Bayesian generative approaches to cluster acoustic segments. Deep learning has been used to learn robust feature representation for speech over varying speaker and background characteristics such as in [13, 14, 15, 16]. [2, 3, 4, 17, 18, 19, 20] explored unsupervised learning of speech features using visual context. Cross-lingual translation research has focused on text-to-text translation [21, 22] as well as speech-to-text from one language to another [23, 24, 25]. [5] recently showed that joint image and speech training performs well on cross lingual caption retrieval using English and Hindi, serving as a basis for speech to speech pseudo translation and [26] confirmed this result using an English-Japanese dataset. A similar line of work was presented in [27], which explored cross-lingual keyword spotting using a visual tagging system.

## 3. Dataset and visual speech model

### 3.1. Dataset

We use the same English-Hindi Places Audio dataset in [5]. This dataset is comprised of natural images sampled from the Places 205 image dataset [28], paired with spoken audio captions describing the content of the images. The English speech captions

http://dx.doi.org/10.21437/Interspeech.2019-1718

were collected via Amazon Mechanical Turk, and the collection process is described in [2]. The collection process for the Hindi speech captions, also collected via Mechanical Turk, is described in [5]. The training dataset consists of 84,480 data triplets, and 1000 validation triplets, with each triplet consisting of an image, and English and Hindi spoken captions.

### 3.2. Model

The model consists of three neural networks - one pretrained VGG16 network [29] for learning image representation and two instances of DAVEnet (audio CNN in [3]) for the audio captions. In this paper, we used an ImageNet-pretrained VGG network finetuned on 400K places images as in [4] and all networks used 1024 final embedding dimension. Training images are augmented using random resized crops to 224x224 and then mean and variance normalized per channel using off-the-shelf Imagenet RGB statistics. The speech waveform is represented by a set of 40 log Mel filterbank energies per 25 ms frame of the speech caption at 10 ms shifts. We train the network for 90 epochs with the 6-way triplet loss H↔E↔I↔H from [5]. We use SGD with batch size 128, and initial learning rate of 0.001 decayed by a factor of 10 every 30 epochs. The model achieved the following recall at 10 scores for image(I), English(E) and Hindi(H) retrieval tasks: E→I (0.571), I→E (0.545), H→I (0.404), I→H (0.393), E→H (0.192), H→E (0.211) .

## 4. Bilingual word discovery

In [5], the authors showed that a visually-grounded model of speech trained to associate both English and Hindi spoken captions with semantically-related images was capable of performing semantic speech retrieval between captions in both languages. Preliminary experiments in that paper suggested that the output feature maps of the English and Hindi speech models could be used to approximately align the segments of both speech signals which referred to the same image region. Our goal in this paper is to automatically extract and cluster these segments into word-like units, and establish pairwise linkages between English and Hindi clusters that capture similar semantics. This section describes the steps we took to discover the word clusters and establish linkages.

### 4.1. Selecting regions of interest

Given an input log Mel spectrogram spanning 40 filterbanks across $T$ temporal frames, the output of the DAVEnet audio model will be a feature map with $d$ channels spanning $\frac{T}{8}$ frames. During training, mean pooling is used to compress this feature map into a single $d$ dimensional vector, but when using an already-trained model for word discovery we do not apply this pooling so as to preserve temporal information for the purpose of word localization. Although the output of the DAVEnet model captures semantics, it does so by producing a dense embedded representation; it does not explicitly segment or tokenize the speech signal. In order to identify regions of interest with a high likelihood of containing a meaningful word, we use an approach inspired by the "interval piling" step of the Segmental Dynamic Time Warping (S-DTW) pattern discovery algorithm [6], which identifies regions of an utterance which exhibit high similarity with regions in many other utterances. While the S-DTW algorithm computes similarities in the acoustic observation space, our method instead compares pairs of utterances in the $d$-dimensional multimodal semantic embedding space learned by the DAVEnet model. For some reference utterance containing a given word, other utterances containing the same underlying word can help inform the location of the word. We rely on a set of nearest neighbors of the reference utterance in order to perform word localization. This is done by obtaining the 1024-dimensional mean-pooled DAVEnet output for each utterance in the training set, and finding its $K$ nearest nearest neighbors according to the dot product similarity. For our experiments, we used $K = 100$.

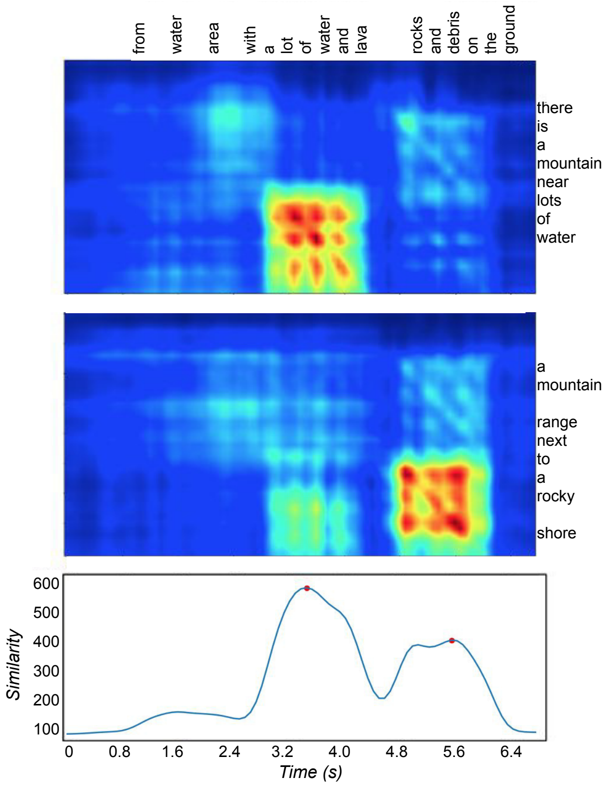


Figure 2: *Embeddings at the speech network output for the top utterance are compared with those from its nearest neighbors. Taking the frame-level maximum over all neighbors results in the similarity profile.*

After selecting the nearest neighbors for each utterance, we compute a set of similarity maps $\{M^1, \ldots, M^K\}$ between the reference and each of its neighbors. To encode temporal localization information in each $M^n$, we extract the outputs of the last convolutional layer of the DAVEnet model before the global average pooling layer. Then, we compute $M^n_{i,j} = R_i \cdot N^n_j$ where $R_i$ represents the $i^{th}$ frame of the DAVEnet output for the reference utterance, and $N^n_j$ represents the $j^{th}$ frame of the DAVEnet output of its $n^{th}$ nearest neighbor. We then compute a similarity profile $p$ where $p_i = \max_{n,j} M^n_{i,j}$. This gives a profile with peaks at locations of the reference utterance that exhibit high similarity to some region of at least one of its neighbors, indicating the presence of a shared word. This can be seen in Figure 2. To facilitate peak picking, we apply a Gaussian smoothing filter to $p$ with $\sigma = 1$. We use comparison among neighboring values to find local maxima and select peaks with prominence of at least $max(200, 0.15 \times r)$ where $r$ is the range of values in $p$.

### 4.2. Clustering regions of interest

The embedding vectors in the locations of the selected peaks represent regions that may contain words whose semantics have been learned by the cross-modal model. All the utterances in the dataset are processed to get these peak locations and their associated embeddings. Next, we cluster the peaks using a Dirichlet Process Gaussian Mixture Model (DPGMM) [30]. Both English and Hindi peak embeddings were normalized together to have zero mean and unit variance and then projected from $\mathbb{R}^{1024}$ to $\mathbb{R}^{300}$ using PCA. The clustering algorithm used kmeans initialization with 200 components, diagonal covariance matrices for each component, mean precision prior of 30, weight concentration prior ($\gamma$) of 1000 and allowed to run for a maximum of 1500 iterations.

### 4.3. Linking English and Hindi clusters

To control for repeated clusters, we put the cluster centroids into matrices $E$ and $H$ for English and Hindi respectively, each with rows representing all centroids in the language. We constructed an undirected graph whose edges connect rows of $E$ with rows of $H$ and edge weights represented by $E_i H_j^T$. Edges with weight less than a threshold $\tau = 300$ were set to zero. Finally, we ran the Louvain graph community clustering algorithm [31] to group the DPGMM English and Hindi cluster centroids into meta-clusters that capture bilingual concepts.

### 4.4. Deriving cluster labels for evaluation

Given $N$ peaks in a cluster, each peak has a receptive field of size $f$. Since the neural network is symmetric for convolution and pooling operations, we simply find the location in the speech caption directly below the peak as $p \times c/n$, where $p$ and $n$ are the location of the peak and the number of frames respectively in the last convolution layer and $c$ is the caption length in seconds. We then snip $f/2$ seconds on both sides of the the selected peak location within the caption. After performing this operation for each peak in the cluster, we select the ASR text transcripts of these portions to present in this paper. To get a single class label, we calculate the purity of each word selected for the cluster. Purity is the proportion of the selected $f$-second windows containing a given word. We also compute coverage, the fraction of the total number of instances of a word in the dataset captured by a cluster. To control for stop words that occur next to salient words, we weight the purity scores by the average duration of the word. We then rank the words by the weighted purity and select the top word as the cluster name. In our experiments, $f$ was approximately 2.5 seconds.

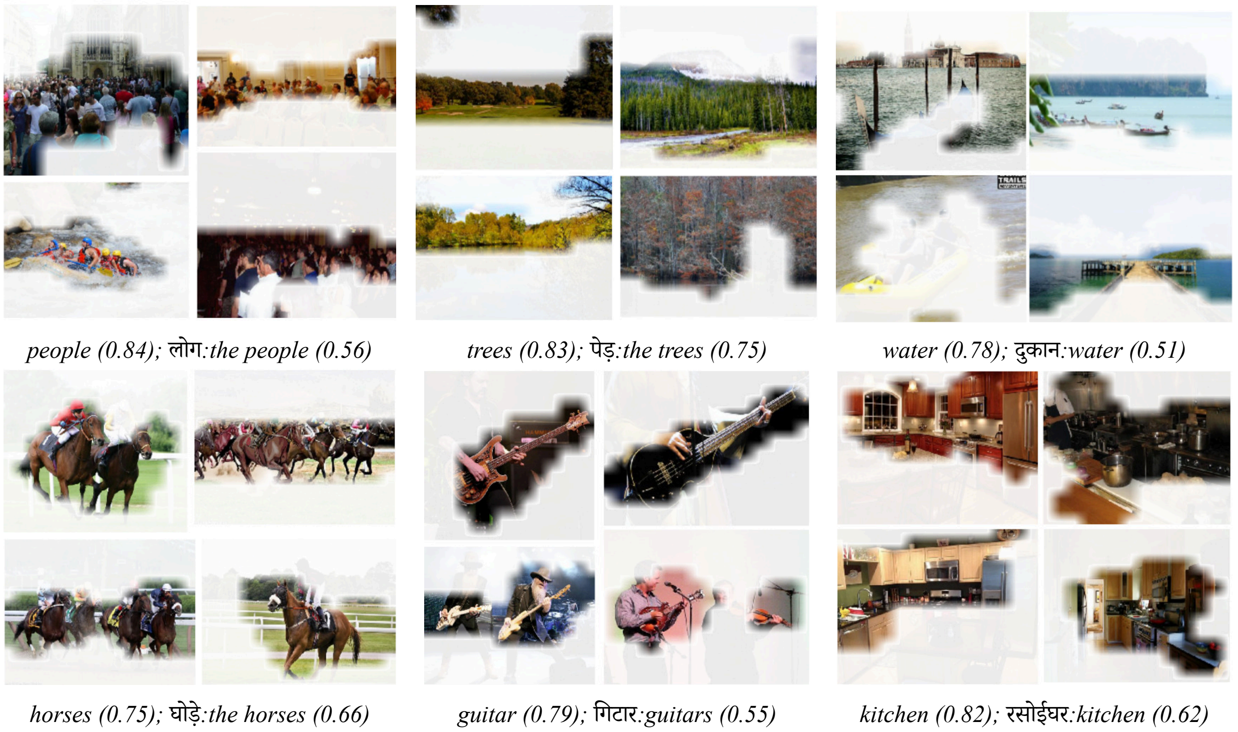
## 5. Experiments

### 5.1. Bilingual word clustering

Details of the top 35 meta-clusters (rank ordered by similarity) out of 101 discovered meta-clusters in the training dataset are presented in Table 1. Each row of the table shows statistics for English and Hindi clusters grouped together in the same meta-cluster. All texts refer to the ASR transcription of the underlying speech, and Hindi texts are paired with their Google Translate API's translation to English. The numbers are reported by merging all English clusters whose centroids exist in a meta-cluster and the same is done for Hindi clusters.

### 5.2. Creating a picture dictionary

To investigate the visual semantics of the bilingual meta-clusters, we find the image regions from the training images



*people (0.84); लोग:the people (0.56)*

*trees (0.83); पेड़:the trees (0.75)*

*water (0.78); दुकान:water (0.51)*

*horses (0.75); घोड़े:the horses (0.66)*

*guitar (0.79); गिटार:guitars (0.55)*

*kitchen (0.82); रसोईघर:kitchen (0.62)*

Figure 3: *Picture dictionary representing three-way agreement between English speech caption, Hindi speech caption and Image pixels. We present the text transcriptions of the clustered speech segments with their corresponding cluster purities.*

Table 1: *Bilingual word clusters.* $E_1$ *and* $E_2$ *correspond to the top two labels for combined English clusters within a meta-cluster and* $H_1$ *and* $H_2$ *are the Hindi equivalents.* $P_E$ *and* $P_H$ *are purity scores while* $C_E$ *and* $C_H$ *are coverage fractions using the top 1 label.* $S$ *represents the similarity score between linked English-Hindi clusters.* $N_E$ *and* $N_H$ *are the number of peaks in English and Hindi respectively in the meta-cluster. Average purity and coverage of the top label across all English clusters are 0.53 and 0.45, while for Hindi they are 0.44 and 0.31. The number of clusters with purity greater than 0.5 is 59 for English and 34 for Hindi.*

| $E_1$ | $P_{E1}$ | $E_2$ | $P_{E2}$ | $H_1$ | $P_{H1}$ | $H_2$ | $P_{H2}$ | $S$ | $N_E$ | $N_H$ | $C_E$ | $C_H$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| lighthouse | 0.55 | house | 0.28 | लाइट:light | 0.36 | हाउस:house | 0.30 | 4898 | 485 | 578 | 0.73 | 0.30 |
| bed | 0.41 | bedroom | 0.29 | बिस्तर:bed | 0.37 | बेड:bed | 0.11 | 2470 | 1620 | 1535 | 0.85 | 0.73 |
| guitar | 0.79 | playing | 0.39 | गिटार:guitars | 0.55 | बजाते:playing | 0.09 | 1961 | 280 | 411 | 0.74 | 0.76 |
| staircase | 0.21 | stairs | 0.23 | सीढ़ियां:stairs | 0.25 | चिड़िया:bird | 0.14 | 1925 | 1115 | 1305 | 0.60 | 0.81 |
| windmill(s) | 0.55 | turbine(s) | 0.14 | पवन:air | 0.47 | चक्की:mill | 0.23 | 1882 | 569 | 627 | 0.69 | 0.74 |
| kitchen | 0.82 | cabinets | 0.02 | रसोईघर:kitchen | 0.36 | रसोई:kitchen | 0.26 | 1827 | 680 | 686 | 0.57 | 0.82 |
| bridge | 0.77 | suspension | 0.04 | पुल:the bridge | 0.27 | ब्रिज:the bridge | 0.23 | 1805 | 1681 | 895 | 0.64 | 0.21 |
| boat(s) | 0.64 | sailboat | 0.04 | नाव:the boat | 0.24 | 'जहाज':ship | 0.04 | 1694 | 1317 | 1229 | 0.60 | 0.76 |
| fish | 0.64 | aquarium | 0.07 | मछली:fish | 0.49 | मछलियां:fish | 0.22 | 1617 | 639 | 467 | 0.63 | 0.57 |
| snow | 0.49 | snowy | 0.13 | बर्फ:ice | 0.47 | बर्फीले:snowy | 0.11 | 1586 | 3464 | 3221 | 0.70 | 0.55 |
| closet(s) | 0.63 | cabinet(s) | 0.01 | अलमारी:cupboard | 0.51 | अलमारियां:shelves | 0.11 | 1537 | 406 | 752 | 0.63 | 0.49 |
| horse(s) | 0.75 | riding | 0.08 | घोड़े:the horse | 0.66 | सवार:the rider | 0.05 | 1431 | 674 | 442 | 0.62 | 0.59 |
| track | 0.37 | running | 0.37 | दौड़:the race | 0.31 | प्रतियोगिता:contest | 0.13 | 1422 | 407 | 408 | 0.25 | 0.43 |
| bus | 0.55 | inside | 0.15 | बस:bus | 0.40 | हवाई:airy | 0.24 | 1319 | 851 | 930 | 0.61 | 0.31 |
| station | 0.52 | subway | 0.38 | स्टेशन:station | 0.48 | रेलवे:railway | 0.45 | 1296 | 471 | 262 | 0.28 | 0.28 |
| church | 0.64 | cathedral | 0.12 | गिरजाघर:the cathedral | 0.27 | चर्च:the church | 0.20 | 1213 | 746 | 687 | 0.42 | 0.74 |
| bird(s) | 0.69 | flying | 0.07 | पक्षी:the bird | 0.35 | चिड़िया:bird | 0.27 | 1171 | 261 | 378 | 0.42 | 0.54 |
| blue | 0.70 | sky | 0.06 | नीले:blue | 0.56 | रंग:colour | 0.65 | 1167 | 2242 | 809 | 0.29 | 0.26 |
| shower | 0.55 | bathroom | 0.24 | नहाने:bathing | 0.30 | बाथरूम:bathroom | 0.14 | 1137 | 583 | 454 | 0.55 | 0.66 |
| man | 0.76 | standing | 0.01 | आदमी:man | 0.41 | व्यक्ति:person | 0.11 | 1068 | 2803 | 3435 | 0.35 | 0.26 |
| mountain(s) | 0.74 | range | 0.03 | पहाड़:the mountain | 0.51 | पहाड़ियों:the hills | 0.20 | 1053 | 888 | 1469 | 0.24 | 0.21 |
| wooden | 0.49 | wood | 0.17 | लकड़ी:the wood | 0.86 | एक:one | 0.24 | 1017 | 1858 | 1007 | 0.42 | 0.36 |
| table(s) | 0.76 | wooden | 0.05 | टेबल:table | 0.24 | मेज:the table | 0.16 | 994 | 2418 | 2191 | 0.42 | 0.22 |
| children | 0.57 | kids | 0.08 | बच्चे:children | 0.60 | बच्चों:the children | 0.16 | 982 | 723 | 1406 | 0.52 | 0.50 |
| people | 0.89 | standing | 0.09 | लोग:the people | 0.69 | कुछ:some | 0.73 | 943 | 2683 | 916 | 0.31 | 0.27 |
| forest | 0.51 | trees | 0.10 | जंगल:forest | 0.48 | है:is | 0.30 | 868 | 821 | 1198 | 0.52 | 0.52 |
| sitting | 0.63 | people | 0.50 | बैठे:sitting | 0.64 | हैं:are there | 0.53 | 862 | 1509 | 1183 | 0.26 | 0.33 |
| child | 0.31 | boy | 0.27 | बच्चा:child | 0.34 | एक:one | 0.52 | 858 | 737 | 611 | 0.41 | 0.41 |
| microphone | 0.43 | stage | 0.14 | मंच:forum | 0.30 | माइक:mike | 0.27 | 841 | 445 | 327 | 0.62 | 0.30 |
| clouds | 0.52 | sky | 0.20 | बादल:cloud | 0.75 | आसमान:sky | 0.42 | 801 | 1002 | 928 | 0.47 | 0.42 |
| water | 0.78 | body | 0.45 | पानी:water | 0.51 | नदी:river | 0.23 | 779 | 1838 | 2828 | 0.22 | 0.27 |
| trees | 0.83 | pine | 0.04 | पेड़:the trees | 0.49 | पेड़ों:the trees | 0.26 | 770 | 1634 | 1396 | 0.24 | 0.13 |
| grass | 0.57 | green | 0.25 | घास:grass | 0.36 | है:is | 0.32 | 728 | 1273 | 2306 | 0.32 | 0.40 |
| wall(s) | 0.73 | on | 0.40 | दीवार:wall | 0.64 | पर:on | 0.47 | 616 | 1431 | 873 | 0.33 | 0.23 |
| yellow | 0.72 | and | 0.20 | पीले:yellow | 0.51 | रंग:colour | 0.52 | 586 | 1220 | 870 | 0.36 | 0.31 |

most similar to each meta-cluster. We use the mean pooled DAVEnet output of the image branch to represent the images and find the top K images that have the highest average (dot product) similarity to all of the DPGMM cluster centroids captured by the meta-cluster. We then compute a binary mask $S$ for each image by taking the inner product of each superpixel region of the DAVEnet image model's output feature map (before global average pooling) with the average of the meta-cluster centroids. We threshold $S$ such that pixels with similarity greater than $.25 * max(S)$ are set to 1 and 0 otherwise. Finally, we upsample $S$ to the same resolution as the original image using bilinear interpolation, and apply the mask to the image. This results in a audio-visual picture dictionary, from which we show several examples in Figure 3.

## 6. Conclusion

We presented a method for discovering bilingual word clusters using a visually grounded model of speech audio. We presented numerical clustering results as well as an audio-visual picture dictionary, demonstrating that our method is capable of discovering clusters of word-like units in both English and Hindi that exhibit a high degree of semantic agreement. In our future work, we plan to develop improved methods for learning a larger number of concepts, especially actions and verbs. Our current approach required us to utilize spoken captions for a common set of images for both languages, but we plan to investigate whether similar results can be achieved when different sets of images are used for each language's captions. We would also like to extend our method from the bilingual case to the multi-lingual case. Future work should also investigate direct speech-to-speech translation using our discovered meta-clusters. Finally, we believe that the representations learned by our acoustic models could find use in traditional ASR systems, such as in low-resource or code switching scenarios.

## 7. Acknowledgements

This research was supported in part by NTT Corporation.

## 8. References

[1] B. Zhou, "Statistical machine translation for speech: A perspective on structures, learning, and decoding," *Proceedings of the*

*IEEE Special Issue on Speech Information Processing*, vol. 101, no. 5, pp. 1180–1202, 2013.

[2] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Proc. Neural Information Processing Systems (NeurIPS)*, 2016, pp. 1858–1866.

[3] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 506–517.

[4] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 649–665.

[5] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4969–4973.

[6] A. Park and J. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[7] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2010, pp. 1676–1679.

[8] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, 2010, pp. 460–470.

[9] D. F. Harwath, T. J. Hazen, and J. R. Glass, "Zero resource spoken audio corpus analysis," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8555–8559.

[10] C.-y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012, pp. 40–49.

[11] L. Ondel, L. Burget, and J. Černockỳ, "Variational inference for acoustic unit discovery," *Procedia Computer Science*, vol. 81, pp. 80–86, 2016.

[12] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 4, pp. 669–679, 2016.

[13] Y. Zhang, R. Salakhutdinov, H.-A. Chang, and J. Glass, "Resource configurable spoken query detection using deep boltzmann machines," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 5161–5164.

[14] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2015, pp. 3199—3203.

[15] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5818–5822.

[16] R. Thiollière, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling." in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2015, pp. 3179–3183.

[17] G. Synnaeve, M. Versteegh, and E. Dupoux, "Learning words from images and speech," in *Proc. Neural Information Processing Systems (NeurIPS)*, 2014.

[18] D. Roy and A. Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science*, vol. 26, pp. 113–146, 2002.

[19] H. Kamper, S. Settle, G. Shakhnarovich, and K. Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017, pp. 3677–3681.

[20] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 27, no. 1, pp. 89–98, 2019.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[22] L. Specia, S. Frank, K. Sima'an, and D. Elliott, "A shared task on multimodal machine translation and crosslingual image description," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 2016, pp. 543–553.

[23] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2625–2629.

[24] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016, pp. 949–959.

[25] S. Bansal, H. Kamper, A. Lopez, and S. Goldwater, "Towards speech-to-text translation without speech recognition," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 474–479.

[26] W. Havard, J. Chevrot, and L. Besacier, "Models of visually grounded speech signal pay attention to nouns: a bilingual experiment on english and japanese," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8618–8622.

[27] H. Kamper and M. Roth, "Visually grounded cross-lingual keyword spotting in speech," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018, pp. 248–252.

[28] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Neural Information Processing Systems (NeurIPS)*, 2014, pp. 487–495.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[30] D. M. Blei, M. I. Jordan *et al.*, "Variational inference for dirichlet process mixtures," *Bayesian analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[31] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.