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.

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
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 feature representation is obtained 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 $H \rightarrow E \rightarrow H$ 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 imaging rate of 0.001 decayed by a factor of 10 every 30 epochs.

Given an input log Mel spectrogram spanning $d$ filterbanks across $T$ temporal frames, the output of the DAVEnet audio model will be a feature map with $d$ channels spanning $T$ frames. During training, mean pooling is used to compress this feature map into a single $d$ dimensional vector. 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 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 neighbors according to the dot product similarity. For our experiments, we used $K = 100$.

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 pooling layer. Then, we compute $M^{n}_{ij} = R_{i} \cdot N_{j}^{n}$ where $R_{i}$ represents the $i^{th}$ frame of the DAVEnet output for the reference utterance, and $N_{j}^{n}$ represents the $j^{th}$ frame of the DAVEnet output of its $n^{th}$ nearest neighbor. We then compute a similarity profile $p$ where $p_{j} = \max_{n} M^{n}_{ij}$. This gives a profile with peaks at locations of the reference utterance which 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 $p$.

![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.](image)
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 \( R^{1024} \) to \( R^{300} \) 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 \( E_{ij} H_{ij}^T \). 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

![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.](image-url)
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 each meta-cluster. We then compute a binary mask such that pixels with similarity greater than $S$ 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, 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
