Towards Multilingual Lexicon Discovery From Visually Grounded Speech

by

Emmanuel Azuh Mensah

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author .....
Department of Electrical Engineering and Computer Science
August 21, 2019

Certified by .....
James R. Glass
Senior Research Scientist
Thesis Supervisor

Certified by .....
David Harwath
Research Scientist
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Towards Multilingual Lexicon Discovery From Visually Grounded Speech

by

Emmanuel Azuh Mensah

Submitted to the Department of Electrical Engineering and Computer Science
on August 21, 2019, in partial fulfillment of the
requirements for the degree of
Master of Engineering

Abstract

In this thesis, 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. Then, 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. Finally, we show the same analysis for a joint training using three languages at the same time, with Japanese as the third language.

Thesis Supervisor: James R. Glass
Title: Senior Research Scientist

Thesis Supervisor: David Harwath
Title: Research Scientist

Acknowledgments

I would like to thank my thesis advisors David Harwath and James Glass, without whose support and guidance this thesis work would not have been possilbe. Many thanks to the Spoken Language Systems group for the support and for providing access to resources to use for my work. Furthermore, my gratitude to the NTT Corporation for their support for my RAship as well as their provision of the Japanese dataset I used in this Thesis. I also appreciate the opportunity to sit in some of the multimodal research meetings between speech and vision groups as these gave me a good example of how fruitful discussions towards the future of Artificial Intelligence looks like. Finally, I want to thank my family and friends for their words of encouragement and support throughout my academic journey.

Contents

1 Introduction **19**

1.1 Motivation 19

1.2 Problem Description 20

1.3 Contributions 21

1.4 Outline 22

2 Related Work **23**

2.1 Unsupervised Speech Processing 23

2.2 Vision and Language 25

2.2.1 Multimodal Learning 25

2.2.2 Cross-Lingual Cross-Modal Learning 27

2.3 Machine Translation 28

3 Background **29**

3.1 Datasets and Data Collection 29

3.2 Speech Signal Representation 31

3.3 Image Representation 32

3.4 DAVEnet Model Training 33

3.4.1 DAVEnet and VGG16 33

3.4.2 ResDAVEnet and Resnet50 34

3.4.3 Training 35

4 Cross-Lingual Keyword Spotting **37**

4.1 Posteriogram Approach 37
4.1.1 Computing Query Word Embedding Vectors 38
4.1.2 Computing Seed 39
4.1.3 Computing Posteriograms 40
4.1.4 Evaluation Method 41
4.1.5 Posteriograms 42
4.2 Keyword Spotting Experiments 43
4.2.1 Results 43
4.3 Keyword Spotting Summary 46

5 Bilingual Lexicon Discovery From Visually Grounded Speech 47
5.1 Method For Bilingual Lexicon Discovery 47
5.1.1 Selecting Regions of Interest 48
5.1.2 Clustering Regions of Interest 51
5.1.3 Linking English and Hindi clusters 51
5.1.4 Deriving cluster labels for evaluation 51
5.2 Experiments 52
5.2.1 Setup 52
5.2.2 Results 55
5.3 Bilingual Lexicon Discovery Summary 57

6 Trilingual Lexicon Discovery 59
6.1 Dataset Extension 59
6.1.1 Data Statistics 60
6.2 Model Extension 60
6.3 Clustering Extension 61
6.4 Experiments 63
6.4.1 Setup 63
6.4.2 Comparing Monolingual, Bilingual and Trilingual Cases . . . 64
6.4.3 Cross-Lingual Word Linkage 67
6.4.4 Picture Dictionary 68

6.5 Trilingual Lexicon Discovery Summary 70

7 Conclusion **73**

7.1 Summary of Contributions 73

7.2 Future Work 73

7.3 Parting Thoughts 74

A Tables **75**

B Figures **95**

List of Figures

3-1 Two examples of quadruplet datapoints, each consisting of an image and ASR transcripts of English, Hindi and Japanese captions. 30

3-2 The DAVEnet model architecture is shown on the upper left as presented in [20]. Conv layers are in blue, pooling layers shown in red, and BatchNorm layer shown in black. Each conv layer is followed by a ReLU. The first conv layer of the audio network uses filters that are 1 frame wide and span the entire frequency axis; subsequent layers of the audio network are hence 1-D convolutions with respective widths of 11, 17, 17, and 17. All maxpool operations in the audio network are 1-D along the time axis with a width of 3. An example matchmap output is shown to the right, displaying a 3-D density of spatio-temporal similarity 34

3-3 The ResDAVEnet model architecture is shown on the upper left. The image branch is based on the ResNet50 architecture, while the audio branch shown is the audio architecture of the ResDAVENet model as presented in [21]. Red blocks represent convolutional layers, gray-blocks indicate BatchNorm layers, yellow block MaxPooling layers, and purple blocks ReLU activations. The four blueblocks in the image branch represent the four bottleneck residual blocks in the ResNet50 model, while the four green blocksin the speech branch represent ResDAVEnet blocks. A schematic diagram of a single ResDAVEnet block is shown in thebottom half of the figure. On the upper right is a sample matchmap output, displaying a 3-D density of spatio-temporal similarity 35

4-1 Matchmap between an English caption (on the right) and Hindi caption (on top). The English translation of the Hindi caption is presented for the top caption. 39

4-2 Frame to Gaussian assignment process. Given a new embedding vector corresponding to a word (eg. Cathedral), we find the posterior probability of the source word being associated with any given cluster (Gaussian), with each cluster containing semantically similar words from both languages. The word is assigned to the cluster with highest posterior probability given the new embedding. 40

4-3 The ROC Plot for *chair* over 10,000 examples. The area below the orange curve is the AUC. 41

4-4 An example of a posteriogram for assigning frames to clusters. The words on the right refer to the cluster names and the words at the bottom are the caption words 42

5-1 Diagram representing our proposed method for multilingual spoken lexi-con discovery. 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). . . . 48

5-2 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. . . 49

5-3 Picture dictionary representing three-way agreement between English speech caption, Hindi speech caption and Image pixels using DAVEnet. We present the text transcriptions of the clustered speech segments with their corresponding cluster purities. 58

6-1 Histograms of caption durations by language. Japanese captions were on average longer than those of English and Hindi, with the Japanese histogram having a longer tail. 61

6-2 Number of words (y axis) which occur log number of times (x axis) in the dataset by language. Japanese captions had more unique words, which occurred more frequently than those of English and Hindi, and had more very high frequency words as well. 62

6-3 Picture dictionary representing four-way agreement between English speech caption, Hindi speech caption, Japanese speech caption and Image pixels using ResDAVEnet. We present the text transcriptions of the clustered speech segments with their corresponding cluster purities. 70

B-1 Picture dictionary representing four-way agreement between English speech caption, Hindi speech caption, Japanese speech caption and Image pixels. We present the text transcriptions of the clustered speech segments with their corresponding cluster purities. 95

List of Tables

3.1 Recall @ 10 scores obtained from training DAVEnet and ResDAVEnet in the bilingual English-Hindi training 36

4.1 Top 5 AUC of English/English retrieval using 200 cluster centers . . 44

4.2 Bottom 5 AUC of English/English retrieval using 200 cluster centers 44

4.3 Top 5 AUC of English/English retrieval using seed words 44

4.4 Bottom 5 AUC of English/English retrieval using seed words 44

4.5 Top 5 ROC of English/Hindi retrieval using seed words 45

4.6 Bottom 5 ROC of English/Hindi retrieval using seed words 45

4.7 English/Hindi retrieval using *building* as a seed word 45

4.8 Examples of Hindi words found using *building* as a seed word 45

5.1 Sensitivity of Dirichlet Process Gaussian Mixture Model parameters using DAVEnet model. The legend for the column headers is *Covariance Type (CVT)*, *Tolerance (TOL)*, *Mean Precision Prior (MPP)*, *Number of seed clusters (N)*, *Weight Concentration Prior (WCP)*, *Maximum Iterations (MI)*, *Number of clusters found (CLUSTERS)*. The last three columns refer to the number of clusters with F_1 Score above 0.5, 0.4 and 0.3 respectively. 54

5.2 Sensitivity of Dirichlet Process Gaussian Mixture Model parameters using the ResDAVEnet model. The legend for the column headers is *Covariance Type (CVT)*, *Tolerance (TOL)*, *Mean Precision Prior (MPP)*, *Number of seed clusters (N)*, *Weight Concentration Prior (WCP)*, *Maximum Iterations (MI)*, *Number of clusters found (CLUSTERS)*. The last three columns refer to the number of clusters with F_1 Score above 0.5, 0.4 and 0.3 respectively. 54

5.3 Sensitivity of meta clustering to edge weight thresholding. N_E is the number of meta clusters with English words, N_H is the number of meta clusters with Hindi words, NU_E and NU_H refer to the number of unique cluster names in English and Hindi respectively. N_T is the number of translations (meta clusters with both English and Hindi). $>.5$, $>.4$ and $>.3$ refer to the number of clusters within the language with F_1 Score above 0.5, 0.4 and 0.3 respectively using the best performing DPGMM result. 55

5.4 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. 56

5.5 Concepts found within clusters presented for three clusters. N refers to the number of instances of the word present anywhere in speech segments found in the given cluster. 57

6.1 Statistics of audio caption lengths in seconds for each language. . . . 60

6.2 Comparing retrieval scores for Hindi and Japanese in the monolingual, bilingual and trilingual settings. I, H and J refer to Image, Hindi and Japanese respectively and the right arrow shows that the item on the left is used to retrieve the item on the right. 65
6.3 Comparing Number of words found for Hindi and Japanese in the monolingual, bilingual and trilingual settings as well as number of clusters with F1 scores greater than three different thresholds. 66
6.4 Comparing retrieval scores for English in the monolingual, bilingual and trilingual settings. 67
6.5 Comparing Number of words found for English in the monolingual, bilingual and trilingual settings as well as number of clusters with F1 scores greater than three different thresholds. 67
6.6 Recalls in the trilingual setting with length constrained input captions. 68
6.7 Trilingual Lexicon Discovery. This table presents the top 3 cluster labels for some of the discovered trilingual clusters. p1, p2 and p3 refer to the purity scores using the top 3 labels respectively and N is the number of speech segments in each cluster. 69
6.8 Statistics of number of occurrences of words for the top labels in the dataset for each language 70

A.1 Bilingual word clusters with each block representing English and Hindi clusters within a meta-cluster. $p1$, $p2$ and $p3$ refer to the purity scores of the top three words in the cluster. N is the number of peaks in the meta-cluster for each language. 75
A.2 Trilingual word clusters with each block representing English, Hindi and Japanese clusters within a meta-cluster. $p1$, $p2$ and $p3$ refer to the purity scores of the top three words in the cluster. N is the number of peaks in the meta-cluster for each language. 86

Chapter 1

Introduction

1.1 Motivation

With the many languages in the world, people often need to cross language barriers to communicate. This need is ever increasing with the globalization of economies, democratization of education, exchange of information across the internet, etc. However, different languages have different amounts of resources available for building technologies such as automatic speech recognition and machine translation, which require large amounts of data. Some of the most promising emerging markets exist in parts of the world where linguistic resources are scarce. High-resource languages like Latin derivative languages have traditionally supported state-of-the-art results in spoken language technologies, thus attracting the most attention in the research community.

Current speech-to-speech translation systems rely on a cascade of models that perform in order, automatic speech recognition, machine translation, and text-to-speech synthesis [51]. These models each require large quantities of manually-annotated training data by linguistic experts, 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 thesis, 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 [22, 19, 20, 18], 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 automatic speech recognition. This type of model was generalized to handle speech inputs from two different languages in [18], and was shown to be capable of cross-lingual matching of semantically similar captions.

1.2 Problem Description

Although speech-to-speech translation without the use of text transcripts is a challenging task, we tackle a crucial first step. We wish to automatically create a pseudo-word speech segment dictionary between languages without the use of text transcripts. This step is important because the ability to find speech segments representing words automatically from raw audio has been explored in [39, 34, 41] but word-level alignment of semantically similar speech segments across languages without supervision has so far not been widely studied.

Given a set of speech captions from two languages, we wish to first learn salient words from each language separately without prior annotation of which words exist or where in the speech captions the words exist. [18] shows that the caption can be represented in an embedding space such that we can compute semantic similarities between speech segments with vector operations such as inner products. With this capability, we wish to find captions containing similar concepts in the same language and aligning these concepts at approximately the word level. After finding words in one language, the next task will be to perform the same discovery process to a second language. As noted in [18], the embedding space is able to group concepts from different languages within the same vicinity. We therefore wish to explore similarities between words learned in the first language and the second language. If successful, this will constitute an approximate word level translation between the two languages based on visual semantics, without the need for text transcripts.

Following the bilingual setting, a natural question is whether the method generalizes to more than two languages at once. We would like to explore possible benefits achievable by using more languages with the same training process as in the bilingual case. Quantitatively, we evaluate performance using a cross-modal and cross-lingual retrieval task as well as the number of concepts (words) learned with the addition of more languages.

Finally, with image region salience presented in [20], we want to confirm that we can automatically select regions of images described by words from the three languages that have been identified as approximate translations of each other. This will form the basis of a picture dictionary, which present multiple concepts in multiple languages along with regions of images they describe.

1.3 Contributions

We contribute towards multilingual lexical discovery from visually grounded speech in the four ways:

1. We present two different methods for automatically discovering a bilingual lexicon from visually grounded speech.

2. Using one of the two methods, we successfully present a bilingual word level translation table. The table shows quantitative descriptions about the strength of the model to identify words in a single language as well as a proxy for measuring the confidence of the translation without the need for text intermediates.

3. Following the bilingual setting, we are able to produce a trilingual translation table using the same process as in the bilingual setting, with minimal modifications to accommodate a third language.

4. Finally, we present a trilingual picture dictionary representing concepts learned by at least one of the three languages and the corresponding regions of images to which they refer.

1.4 Outline

The remainder of this thesis is organized as follows. We first discuss related work to unsupervised speech processing and machine translation as well as new multimodal speech processing methods that combine speech and vision in Chapter 2. Chapter 3 discusses the models, datasets and preprocessing used in the rest of the thesis. Then in Chapter 4, we present our initial attempt at automatic lexicon discovery and preliminary results. Chapter 5 presents our final work on the bilingual lexicon discovery, following which we show the extension in the trilingual case, as well as presenting a picture dictionary in Chapter 6. Finally, we discuss possible follow up work to this project in Chapter 7. Full details of the translation tables and picture dictionary are presented in appendices of this thesis.

Chapter 2

Related Work

Our work stands at the intersection of unsupervised speech processing, joint vision and language learning and machine translation. Each of the fields has seen large amounts of progress and we build off several techniques identified by the various research threads. In this chapter, we review the related work to this Thesis.

2.1 Unsupervised Speech Processing

Current supervised methods of speech recognition and machine translation can be prohibitively expensive, especially due to the need for large volumes of data and high quality transcripts of conversational speech [37]. This high cost has prompted research into methods for increasingly unsupervised speech recognition, which we describe in this section.

A landmark unsupervised speech processing algorithm uses Segmental Dynamic Time Warping (S-DTW) proposed by [39] to automatically find patterns within speech to discover lexical entities directly from untranscribed audio stream. The algorithm is able to find repeated regions of high acoustic similarity, allowing the discovered segmentations to be grouped into linguistic units such as words or phrases. This paradigm represented a shift from the traditional supervised classification of speech segments into prespecified lexical units. [26] extends S-DTW by using a coarser grained first pass, made possible by the low occurrence frequency of terms, gaining

improvement in computational complexity over the $\mathcal{O}(n^2)$ S-DTW algorithm to make it more applicable at scale. [23] further uses the S-DTW and a probabilistic approach to summarize main topics found in speech corpora, while [12] extends the S-DTW paradigm for natuaral language processing tasks such as document clustering and classification.

Another line of work uses a Bayesian generative approach to cluster acoustic segments. [34] uses a Dirichlet process mixture model where each mixture is a Hidden Markov Model (HMM) representing a sub-word unit. The generative process involves finding a set of sub-word units and HMMs that best explains the data, through an iterative inference process. [38] shows that a Variational Bayesian (VB) inference approach can both be more practical and accurate than Gibbs Sampling used in prior Bayesian work in acoustic unit discovery as VB can be easily parallelized, even though VB tends to find local optima. Other works such as [29] used a Bayesian generative approach to segment raw audio input into possible word segments of arbitrary lengths, maps these segments into a fixed dimension acoustic vector space and clusters the discovered groups.

Deep learning approaches have proved very useful in semi-supervised and unsupervised speech recognition, especially for generating robust feature representation for speech over varying speaker and background characteristics [41]. [49] uses Deep Boltzmann Machines to generate posterior probabilities in a query detection task. The method trained using only under a third of annotated data, is able to perform on par with systems trained on fully labelled datasets. [28] uses an unsupervised approach to train a robust feature extractor in a zero-resource setting for speech by first using autoencoding to initialize a neural network and then using dynamic programming approaches used in Unsupervised Term Discovery, enforces finer feature encoding when similar speech segments are used in a correspondence autoencoder.

2.2 Vision and Language

2.2.1 Multimodal Learning

At the intersection of computer vision, natural language processing and speech processing is a growing interest in the association between vision and language for multimodal learning. These come in two formats: vision-text and vision-speech. Traditionally, multimodal vision and language work has used English text. In contrast, using speech (audio) is a much hard problem due to difficulty of separating continuous speech into word segments, ambiguity in resolving homophones, among several other challenges. An early approach used phoneme strings as a step away from transcribed text in the speech-vision learning task. Recent years have however pushed for increasing research into direct correspondence between vision and audio waveform of speech.

Vision and Text

[6] uses statistical correspondence between segmented image regions and words to be used in tasks such as auto-annotation. In [45], images are segmented and labeled in a semi-supervised learning setting using a small set of labeled images and a large unaligned text corpus by embedding visual and textual words in a latent meaning space. [36] investigates compositional meaning representation using a probabilistic categorical grammar and classifiers for physical characteristics. The Deep Visual-Semantic model introduced by [15] shows that the number of categories recognized by visual models can be significantly increased by adding a non parallel text dataset to both help with training and constrain the predictions. The model is able to use semantic information to make predictions about image labels not observed during training. [35] explores joint visual and textual training in video retrieval tasks using complex natural language queries by creating a semantic graph based on the video description and matching it to visual concepts using a generalized bipartite matching algorithm. In addition to the 2D and video cases for vision, [33] is able to perform semantic parsing in 3D scenes.

Caption generation is another task that has received attention in the vision-language space. [31] uses convolutional neural networks over image regions, recurrent neural networks over text and a multimodal objective that learns alignment between the two modalities to generate new descriptions for image regions. A similar technique is presented in [47], where they maximize the likelihood of a target description sentence given an image to produce natural sentences describing an image. [14] tackles the description generation using a set of proposed descriptions generated by statistical methods and ranking them by sentence level features and a multimodal similarity model using deep learning.

Aside image segmentation and caption generation, many other topics have been explored in the vision-language space. Visual question answering as presented in [2], seeks to provide a natural language answer given an image and a question about the image. [11] also introduces an visual dialog system and [40] synthesizes images from text descriptions.

Vision and Speech

An early novel work in image-speech learning introduced by [42] used phoneme strings to represent speech in the speech-vision learning. This approach relaxed the constraint on purely transcribed speech in multimodal learning. Recent work has proved that image-speech multimodal learning has traction even in the absence of text, by dropping any constraint on pre-processing speech by transcription or phoneme annotation. [22] showed that neural network models could be trained to learn semantic concepts across both visual and spoken modalities without expert linguistic knowledge and these correspondences used in a retrieval task. Further, [19] is able to automatically find spoken instances of words and associate them with objects within images they describe. The speech-image models have been shown to learn linguistic units such as words and phones in their internal representations. [13] shows that features learned in audio-visual models are less speaker dependent than traditional speech recognition approaches and contain more discriminative linguistic information. Audio-Visual deep networks have also been shown to capture more meaningful semantic informa-

tion higher in the hierarchy of layers while form-related parts of language is stronger earlier in the networks [10], and this finding is further stressed in [3] that phoneme representation is most defined in the lower layers of the networks.

2.2.2 Cross-Lingual Cross-Modal Learning

Another line of work that has recently gained attention in the research community is using visual and text or speech modalities for learning correspondences between languages. [7] introduced an implicit method of learning bilingual pairs of words as proposed translations by using monolingual image-to-word connections in labelled web images. They associated texts from two languages if their corresponding images had similar visual features. Using this method, they showed that using visual context improved translation accuracy over string edit distance based approaches. [16] learned a multilingual multimodal representation such that captions from two languages describing the same image would be placed close to each other in the same multimodal embedding space. They did this by introducing a pairwise ranking loss function which could handle both symmetric and asymmetric similarity between the two modalities. This method was shown to achieve state-of-the-art performance in image-description ranking for German and English and textual similarity of image descriptions. [18] recently showed that joint image and speech training performs well on cross lingual spoken caption retrieval using English and Hindi, serving as a basis for speech to speech pseudo translation. [24] also confirmed this result using a English-Japanese dataset. A similar line of work was presented in [30], which explored cross-lingual keyword spotting using a visual tagging system. [43] recently released a multimodal collection of instructional videos with English subtitles and Portuguese translations, along with sequence-to-sequence baselines for machine translation, automatic speech recognition, spoken language translation and multimodal summarization. This dataset is meant to stimulate more research activity in the cross-modal cross-lingual space.

2.3 Machine Translation

Machine translation is a well studied problem but traditionally in the text-to-text case and using statistical approaches. [32] presents benefits of phrase-based statistical approaches over word-level methods by formulating translation probability for translating a sentence in one language into another using Bayes rule after segmenting the sentence into a sequence of phrases. The current paradigm uses neural networks such as [5], which uses an encoder-decoder model to translate text from a source language into text in the target language. Although the majority of machine translation systems work in the text-to-text paradigm, recent work have explored the speech-to-text case. [48] uses an attention based sequence-to-sequence architecture to translate speech audio from a source language directly into text in the target language and this concept is further explored in [4, 46]. Even in this case, text-to-speech post processing is still typically required in a speech-to-speech translation system. A new line of research in machine translation uses visual cues to provide context for unsupervised translation such as presented in [46], where they generate a caption in a target language for an image, given an image and/or one or more descriptions in the source language. [27] recently introduced an attention-based sequence-to-sequence neural network which is able to perform end-to-end translation from speech in one language into speech in another without the need for text transcripts. This is done by mapping spectrograms in one language to spectrograms in the target language. Even though the proposed method slightly underperforms the baseline of a cascade of speech-to-text translation and text-to-speech synthesis model, it shows that the difficult task of direct speech-to-speech translation is feasible.

Chapter 3

Background

In this section, we describe the dataset, pre-processing of the dataset as well as the models and training procedure we used for the rest of the thesis. In the cross-lingual lexicon discovery process, we make use of embedding vectors generated from applying these models to the image/audio dataset.

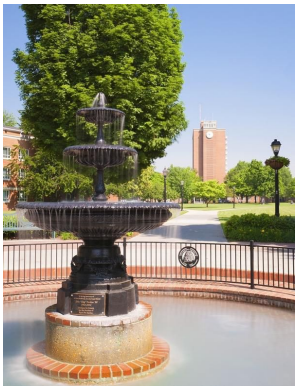
3.1 Datasets and Data Collection

We use the same Places Audio dataset used by [18, 20]. This dataset is comprised of natural images sampled from the Places 205 image dataset [50], 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 more detail in [22]. The Hindi speech captions were also collected via Mechanical Turk, and their collection is described in [18]. The complete English/Hindi training dataset used in our bilingual lexicon discovery experiments consists of 84,480 data triplets, with each triplet consisting of an image, an English spoken caption, and a Hindi spoken caption. We use the same 1,000 triplets validation set as [18]. In the trilingual case, the Japanese caption dataset was collected by the NTT Corporation [TODO: cite] and the overall Trilingual dataset consists of ~74K Image/English/Hindi/Japanese quadruplets and a separate 1K quadruplet for validation. A sample of the datapoints for the trilingual case is shown in Figure 3-1

English
In this photograph there are four people there are two women and two men they all seem to be holding drinks and smiling

Hindi
दो आदमियों और दो औरतों खड़ा हुआ है वह लोग हंस रहा है (Two men and two women are standing, they are laughing)

Japanese
暗い 屋外 か パーティー 会場 の よう な 場所 で あろう か よん 人 の 男女 の 下 の 画像 で ある と 思わ れる 中央 に 賛成 し たり 両脇 に 女性 2人 が 降り 右側 の 犯人 は 家 に 飲みかけ の ドリンク の コップ を 持っ て いる(It might be a dark outdoor place or a place like a party venue. It seems to be the image below the man and woman. In favor of the center, two women got down on both sides and the woman on the right side took a drink to drink at home. Have a cup)



English
This is a picture of a fountain at a park in sea water in the basin of mountain you can see a pathway going and you can see a tall building with a clock on it if it was built in the fact these

Hindi
एक फाउंटेन का नजारा पानी की जिस पर से पानी गिरता हुआ दिखाई दे रहा है और पीछे एक विशाल पेड़ (The view of a fountain of water on which water is falling, and behind it is a huge tree)

Japanese
手前 に さん 段 重ね の オブジェ の よう な 噴水 が 見え て いる その 噴水 の ある 池 は 柵 に 覆わ れ て い て その 先 に は 木 が たくさん 茂っ て いる 公園 の よう に なっ て い て ずっと 奥 に は 茶色い 時計 台 の よう な 建物 が 建っ て いる (You can see a fountain like a multi-layered object in the foreground. There is a building like a brown clock tower)

Figure 3-1: Two examples of quadruplet datapoints, each consisting of an image and ASR transcripts of English, Hindi and Japanese captions.

3.2 Speech Signal Representation

In this section, we detail the pre processing applied to source audio captions before feeding them to any of the audio Convolutional Neural Networks used in this thesis.

The audio waveform recorded by Mechanical Turk workers is digitized using a sampling frequency of 16 kHz, which is sufficient as most of the salient features of the speech signal is contained in the bands below 8 kHz. In Automatic Speech Recogntion (ASR) systems, a Short-Time Fourier Transform (STFT) is typically applied to the discrete signal to produce a good representation of the information obtained from the variations in the vocal tract since a single discrete sample is often not enough to convey the needed information. The STFT is obtained by applying the Fourier Transform independently to windows of the discrete signal, with overlap between successive windows in order to produce smoothly varying frames. In our experiments, we select commonly used parameters in ASR systems which are 25 millisecond windows at 10 millisecond shifts, resulting in an overlap of 15 milliseconds between consecutive windows.

In practice, before performing STFT, the DC component is removed from the discrete signal

$$x_0[n] = x[n] - \frac{1}{N}\sum_{n'=0}^{N} x[n']$$

after which pre-emphasis filtering is applied to flatten the spectrum, balancing the lowpass response from glottal excitation:

$$x_p[n] = x_0[n] - 0.97x_0[n-1]$$

Then, the STFT is computed as

$$X[m,\omega] = \sum_{n=-\infty}^{\infty} x[n]w[n-mR]e^{-j\omega m}$$

where $-\pi \leq \omega \leq \pi$ is the frequency index (ω is half the sampling rate), m is an integer which indexes into the STFT frames, R represents the shift between frames

and w is the window function that selects which samples to include in the window. In this thesis, we used the Hamming window:

$$w[n] = 0.54 - 0.46cos\Big(2\pi\frac{n}{N}\Big), 0 \leq n \leq N$$

where $N = L - 1$ and L is the window length. $X[m, \omega]$ is converted to the power spectrum as the frequency component of $X[m, \omega]$ is deemed to carry little perceptual importance in audition compared with the energy distribution on the frequency axis. The transform from complex to power spectrum is given by:

$$X_p[m, \omega] = |X[m, \omega]|^2$$

Finally, the power spectrum is passed through a series of nonlinearly spaced (along the frequency axis) bandpass filters, known as the Mel scale, to group different frequency components perceived to be linearly the same by humans. The Mel-frequency spectral coefficients, X_{mfsc} are computed as:

$$X_{mfsc}[m, l] = \sum_{k=-\infty}^{\infty} X_p[m, \omega]V_l[k]$$

where V_l is the l^{th} Mel filter and the energies within each Mel filter converted to the dB scale to produce the log Mel-filterbank features:

$$X_{lmf} = 10log_{10}X_{mfsc}[m, l]$$

For all our experiments, X_{lmf} is the spectrogram used to represent the speech audio signal, with 40 MFSCs computed for every 10 milliseconds of speech audio.

3.3 Image Representation

The images are represented by RGB (Red, Green, Blue) channels and each channel is a matrix of pixel values. These pixel values are used as a digital representation of the continuous variation in the intensity of each of the primary colors, with the

pixels taking on floating point values from 0 to 1 (as opposed to integers in the range 0 - 255) . To pre-process images for training, we resize them such that their smaller dimension is 256 pixels. We then take a random crop of 224x224 (but a center crop of the same size during validation). Afterwards, the pixels are normalized to have zero mean and unit variance, according to a global off-the-shelf Imagenet RGB mean and variance (per-color).

3.4 DAVEnet Model Training

In this thesis, we use two types of models in our audio-visual training tasks - DAVEnet (Deep Audio-Visual Embedding Network) and ResDAVEnet (Residual DAVEnet). We describe these models further as well as the procedure we used in training them in this section.

3.4.1 DAVEnet and VGG16

DAVEnet introduced in [20] consists of an image and an audio network, jointly trained to map images and speech captions that are associated with each other to be similar in an embedding space. The image channel of DAVEnet is the VGG16 architecture [44]. Our experiments use an embedding space in $\mathbb{R}^{1024}$ and so we apply a 3x3, 1024 channel linear convolution with no non-linearity to the 14x14 feature map with 512 channels produced by VGG16. The audio channel is a fully convolutional deep neural network which takes as input a spectrogram created as described in Section 3.2. The network has five convolutional layers as described in [19] with a Batch Normalization layer preceding the first layer and the output modified to produce a feature map across the audio rather than a single embedding vector. The DAVEnet architecture is shown in Figure 3-2

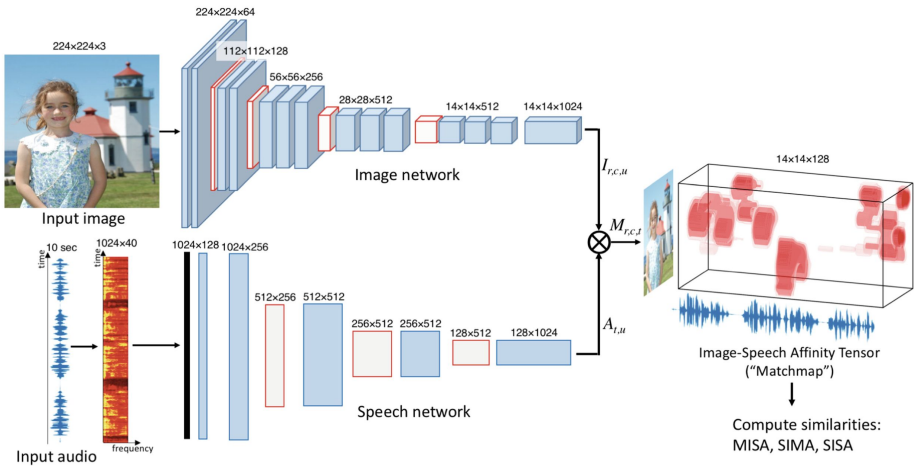


Figure 3-2: The DAVEnet model architecture is shown on the upper left as presented in [20]. Conv layers are in blue, pooling layers shown in red, and BatchNorm layer shown in black. Each conv layer is followed by a ReLU. The first conv layer of the audio network uses filters that are 1 frame wide and span the entire frequency axis; subsequent layers of the audio network are hence 1-D convolutions with respective widths of 11, 17, 17, and 17. All maxpool operations in the audio network are 1-D along the time axis with a width of 3. An example matchmap output is shown to the right, displaying a 3-D density of spatio-temporal similarity

3.4.2 ResDAVEnet and Resnet50

We also use a second pair of model architecture, ResDAVEnet, which consists of two deep residual networks [25]. The image channel of ResDAVEnet uses Resnet50 architecture. The audio channel has a 128 convolutional unit as the first layer spanning one temporal frame but all frequency channels, with a temporal frame shift of 1. The first layer is followed by a ReLU activation layer and a Batch Normalization layer. The remainder of the network consists of four residual stacks with channel dimensions 128, 256, 512 and 1024. Each residual stack has a sequence of two residual blocks as described in [25], which share the same overall channel dimension and the 2D convolution kernels replced by 1-D kernels of length 9. The first residual block of each layer in each stack uses a stride of 2, making the overall network effectively downsample by a factor of 2^4. The ResDAVEnet architecture is shown in Figure 3-3.

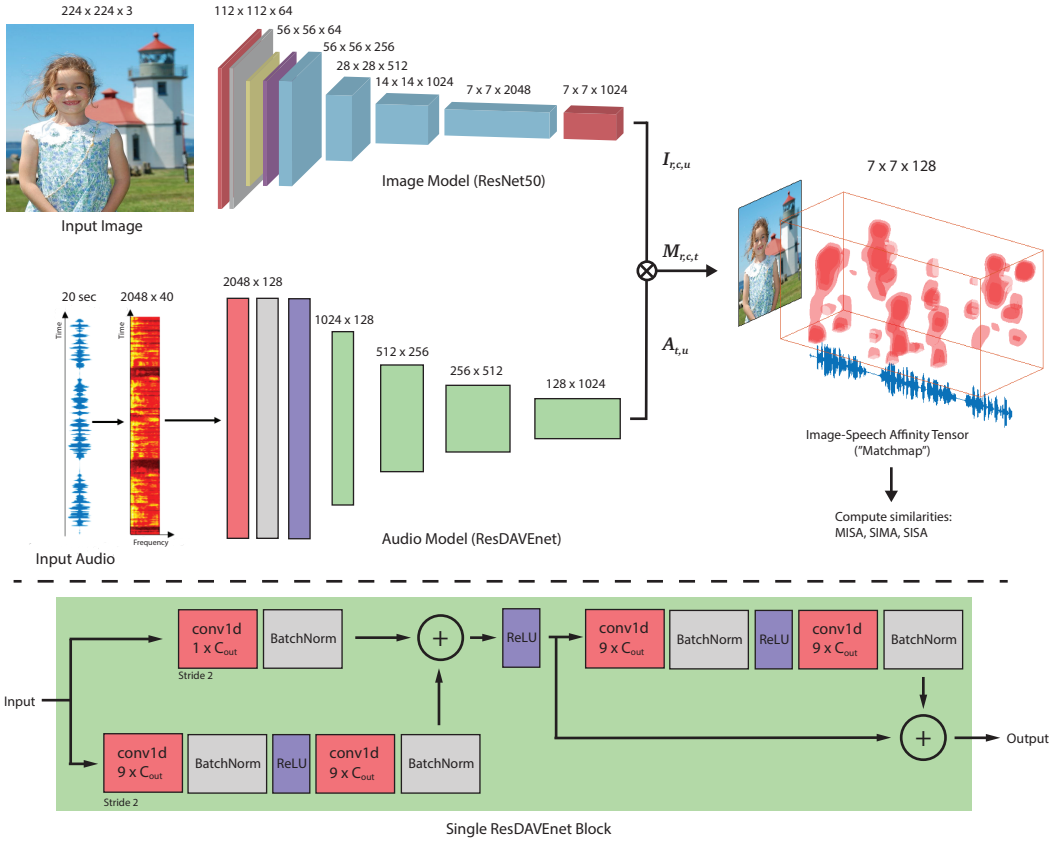


Figure 3-3: The ResDAVEnet model architecture is shown on the upper left. The image branch is based on the ResNet50 architecture, while the audio branch shown is the audio architecture of the ResDAVENet model as presented in [21]. Red blocks represent convolutional layers, grayblocks indicate BatchNorm layers, yellow block MaxPooling layers, and purple blocks ReLU activations. The four blueblocks in the image branch represent the four bottleneck residual blocks in the ResNet50 model, while the four green blocksin the speech branch represent ResDAVEnet blocks. A schematic diagram of a single ResDAVEnet block is shown in thebottom half of the figure. On the upper right is a sample matchmap output, displaying a 3-D density of spatio-temporal similarity

3.4.3 Training

The bilingual training process involves one image model (VGG16 or Resnet50) for learning image representation and two instances of the audio models (DAVEnet or ResDAVEnet audio network) for the audio captions. We train the networks for one round of 90 epochs with the 6-way triplet loss H↔E↔I↔H from [18], with I representing image, E representing English and H representing Hindi. We use a stochastic

gradient descent with batch size of 128, momentum of 0.9 and an initial learning rate of 0.001 decayed by a factor of 10 every 30 epochs.

To initialize the weights of the networks, we first train a monolingual model with ~400K Places Images/English caption dataset, using the same training procedure mentioned above. The pretrained weights are then used to initialize our networks for the cross-lingual training. That is, in the bilingual case, the English and Hindi channels are initialized with the 400K pretrained English network and trained on a separate 84,480 Image/English/Hindi triplet dataset, while the image channel is initialized with the 400K pretrained VGG16 or Resnet50 model. In the trilingual case, all audio networks are initialized with the 400K pretrained English network and trained on a separate 74K Image/English/Hindi/Japanese quadruplet dataset. Refer to Section 6.2 for details of the trilingual extension to the training process as well as recall results for the trilingual case.

We use recall scores to evaluate the performance of the models. Given a set of target vectors $\{v_1, ..., v_n\}$, and a query vector q, if we know that q is associated with one v_i but don't know which, we compute a similarity ($q^T v_i$ in this case) between q and each v_i and rank the result in decreasing order. If the target vector is in the top k, we count a hit and a miss otherwise. We compute this across the evaluation set and refer to the average of the scores as the recall at k. We present the recall at 10 scores for the bilingual model on the validation set of 1000 data triplets in Table 3.1. E, H and I refer to English, Hindi and Image respectively. The letter on the left of the $\rightarrow$ sign is in the domain of q and that on the right, in the domain of v_i

Model	E→I	I→E	H→I	I→H	E→H	H→E
DAVEnet	0.571	0.545	0.404	0.393	0.192	0.211
ResDAVEnet	0.811	0.783	0.609	0.587	0.439	0.449

Table 3.1: Recall @ 10 scores obtained from training DAVEnet and ResDAVEnet in the bilingual English-Hindi training

Chapter 4

Cross-Lingual Keyword Spotting

In this chapter, we present an approach for bilingual lexicon discovery, using a probabilistic model over the embedded representation of speech segments to retrieve occurrences of semantically similar speech segments (at approximately the word level) in both English and Hindi. We used DAVEnet's audio network to generate the speech embeddings. We then compute the area under the Receiver Operator Characteristic (ROC) curve (AUC) [17] to estimate how well our method presents true positive versus false positive matches for translation candidates. In the sections that follow, we discuss the posteriogram approach for finding bilingual word level translations. Finally, we present our results using this approach.

4.1 Posteriogram Approach

In [18], the authors demonstrated 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 underlying image region. Our goal is to automatically extract and cluster these segments into word-like units, and then establish pairwise linkages between English and Hindi clusters that

capture similar semantics. In this section, we present a detailed approach for bilingual lexicon discovery using the posteriogram approach. The process of getting the word level translations is summarized as getting embedding vectors for the source word, computing a posterior probability on a word in the dataset being semantically similar to a source word and then calculating the AUC based on this posterior.

4.1.1 Computing Query Word Embedding Vectors

In order to find other occurrences of a word in our dataset, either in the source or target language, we need a vector against which we will compute similarity to decide whether a new word is close enough to our source word. We do this in one of two ways: automatically via clustering, and in a semi-supervised fashion by leveraging a small amount of aligned text data.

Computing Clusters

To find clusters of similar words, we go through the following process:

1. Find regions of high similarity between English Hindi caption pairs as shown in Figure 4-1. This is done by computing an inner product between the embedding vector of each frame from of one language with the embedding vectors of all frames from the other language. We refer to the resulting similarity matrix as a matchmap and the regions of high similarity as regions of interest.

2. Select the embedding vectors in these regions of high similarity, by running a blob detection algorithm (such as Laplacian of Gaussian filtering) on the matchmaps. A sample of blob selection is shown in Figure 4-1.

3. With the blobs selected, we then perform K-Means clustering in the embedding space to group blobs that align with semantically similar words. In order to figure out the name of the clusters, we compute an intersection-over-union (IoU) score for all words in the cluster. We do this by:
   i. finding the intersection between each word and the regions of high similarity.

ii. find the union between each occurrence of a word in an utterance and a region of high similarity.

We compute this for each word and find the average intersection over union per utterance across all examples in the cluster. The word in the cluster with the highest IoU score in the cluster is used as the cluster name.

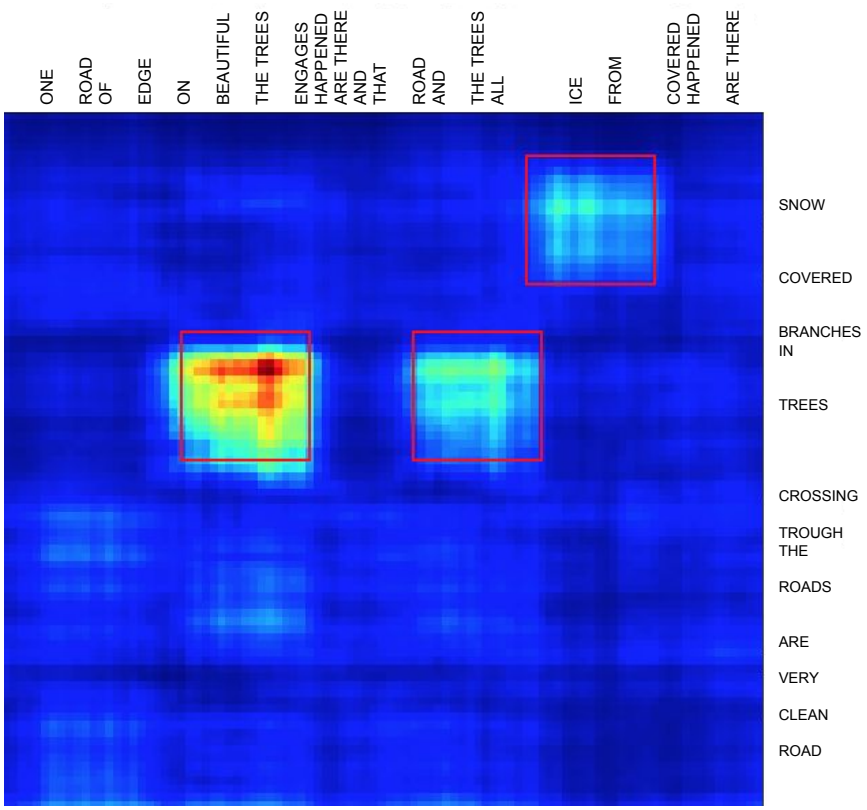


Figure 4-1: Matchmap between an English caption (on the right) and Hindi caption (on top). The English translation of the Hindi caption is presented for the top caption.

4.1.2 Computing Seed

The second method of finding the source vector is a semi-supervised approach where we start with a list of "seed words", chosen beforehand to be the query words. Given these words, we find all occurrences of the words in some fraction of the data for which we assume we have time-aligned text transcripts. We select the embedding vectors from frames corresponding to occurrences of the seed words.

4.1.3 Computing Posteriograms

Given the groups of embedding vectors for a cluster of audio segments representing a word (either seeded or automatically generated), we compute the mean vectors and diagonal covariance matrices for each cluster. Then given a new embedding vector for a new frame in an utterance, we can assign that frame to one of the word clusters based on the posterior probability of the Gaussian g given the frame x according to the equation:

$$P(g|x) = \frac{\mathcal{N}(x; \mu_g, \Sigma_g)}{\sum_0^K \mathcal{N}(x; \mu_{g_k}, \Sigma_{g_k})} \tag{4.1}$$

We call the set of posteriors computed for every frame in a new utterance a posteriogram.

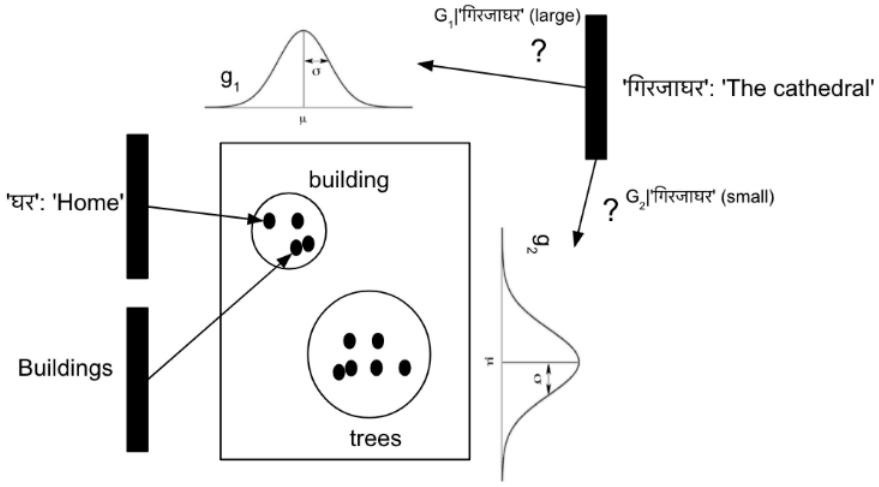


Figure 4-2: Frame to Gaussian assignment process. Given a new embedding vector corresponding to a word (eg. Cathedral), we find the posterior probability of the source word being associated with any given cluster (Gaussian), with each cluster containing semantically similar words from both languages. The word is assigned to the cluster with highest posterior probability given the new embedding.

4.1.4 Evaluation Method

In order to evaluate how well our Gaussians are able to model different occurrences of an underlying word, we formulate a keyword spotting task in which the goal is to search a dataset for utterances that contain an instance of the word in question. We evaluate this task using the receiver operator characteristic (ROC) curve to see the true positive rate against the false positive rates at various thresholds. We report the area under the ROC curve in our experiments. To calculate the ROC for a word, we use the ground truth of whether an example has the word at the utterance level. This is easy because we have the ground truth text caption. The prediction is the posterior probability of assigning any given example to a word group's Gaussian (we use the maximum posterior over all frames in the example). The closer the AUC is to 1, the better our retrieval task is performing. An AUC of 1 is perfect keyword spotting while a score of 0.5 denote a random chance of finding the right example containing the word in question. This example of the ROC plot for *chair* receives a 0.93 AUC over a fourth of the number of examples presented in the results section.

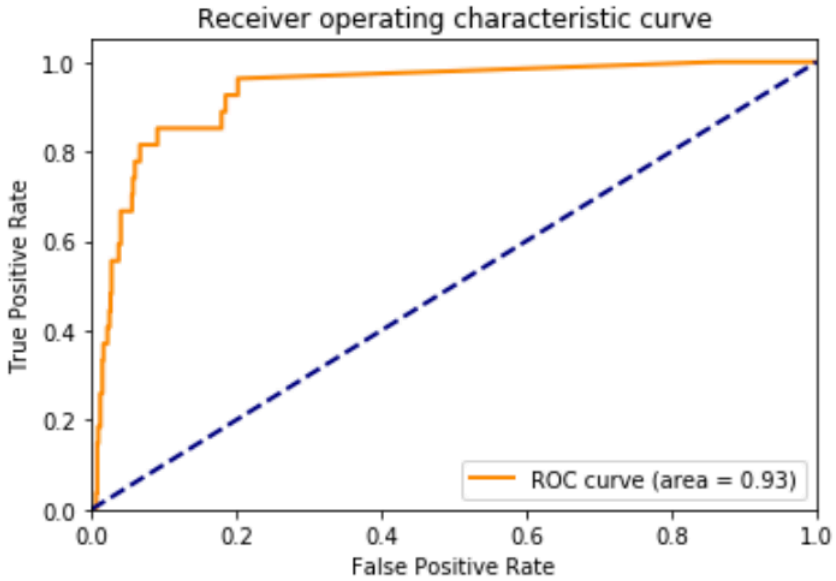


Figure 4-3: The ROC Plot for *chair* over 10,000 examples. The area below the orange curve is the AUC.

4.1.5 Posteriograms

We present a sample visualization of how an example is getting assigned to any of 200 clusters.

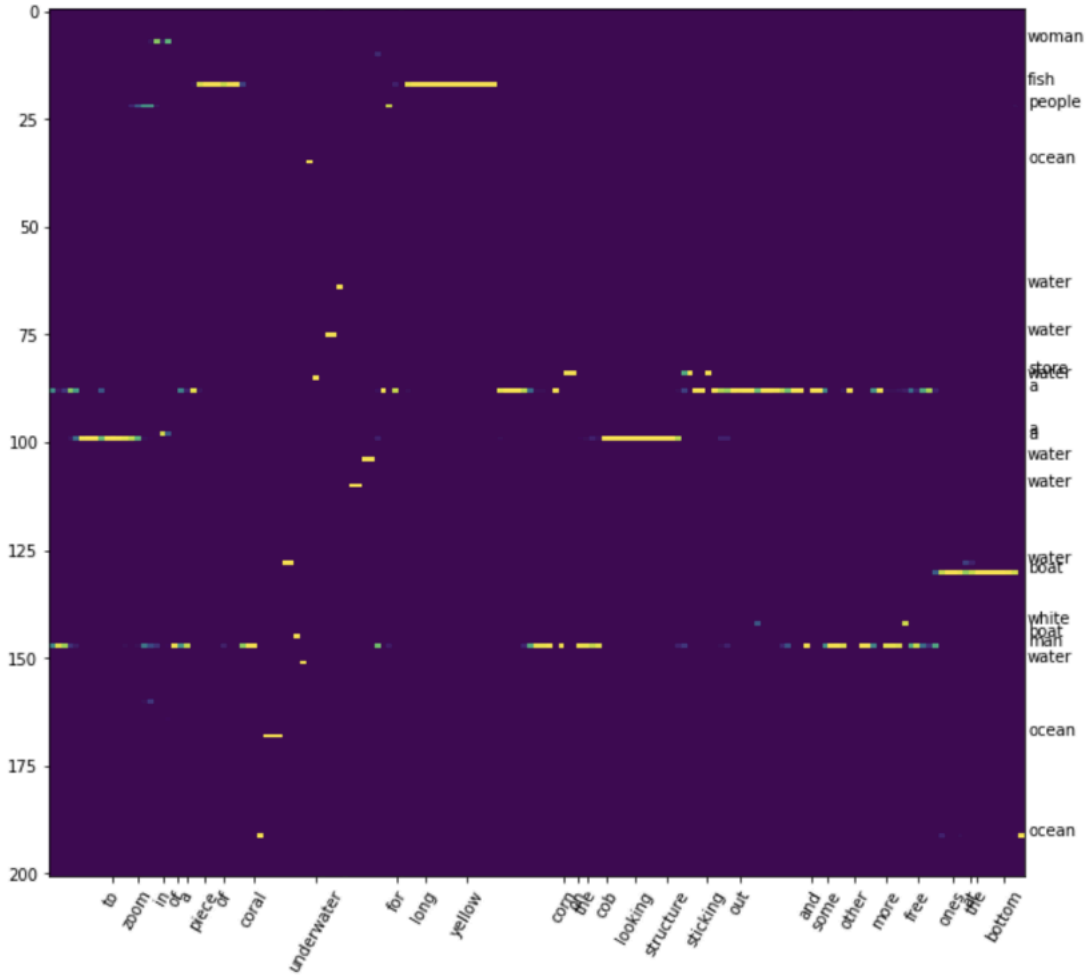


Figure 4-4: An example of a posteriogram for assigning frames to clusters. The words on the right refer to the cluster names and the words at the bottom are the caption words

The caption to the above example is *to zoom in of a piece of coral underwater for long yellow corn on the cob looking structure sticking out and some other more free ones at the bottom.* Here we visualize what the posterior of assigning a frame from an example to any cluster looks like. The pixel values represent the posterior probability that an audio frame in a caption belongs to a particular cluster. The image above has height 200, representing 200 clusters. We only show the words of clusters for which

at least one pixel is above probability of 0.01. We see *coral* associated with *fish* and *ocean* and *water* associated with *underwater*. Words like *people, woman* have small confidences assigned to them.

Finally, we find in our experiments that words like *a* occurred many times, therefore becoming a sort of **background model** where words not immediately assigned strongly to one cluster will be assigned to the background model.

4.2 Keyword Spotting Experiments

Given a query word, our aim is to find occurrences of the query word in English and in Hindi. In this section, we present initial experiments and results for the keyword spotting task. The following sections present initial results of our experiments and brief discussions.

4.2.1 Results

English-English Keyword Spotting

We present how well we assign the right word in the target language to the query word. We show this using the Area Under the Receiver Operator Characteristic curve (AUC). In this proposal, we present the AUC at the utterance level. That is, we count a hit (true positive) if the target word is anywhere in the caption of an example found by a query word. The AUC values presented here are calculated over a total of approximately 40,000 examples.

The table above shows the top 5 and bottom 5 performing of 200 clusters. The *Positives* column presents the ground truth number of examples with the query word in them. Words that occur very frequently but carry little meaning like *a* and *there* end up with low AUC scores. Of the 200 clusters, 96 clusters scored above 0.8 AUC. Next, we present the utterance level AUC of finding occurrences of English frames belonging to seed words. The seed words presented here were selected from the word

Name	Positives	AUC
Skyscapers	82	0.976
Train	605	0.965
Baseball	267	0.962
Snow	1060	0.959
Kitchen	557	0.9957

Table 4.1: Top 5 AUC of English/English retrieval using 200 cluster centers

Name	Positives	AUC
There	9685	0.514
a	31005	0.507
a	31005	0.505
There	9685	0.502
a	31005	0.501

Table 4.2: Bottom 5 AUC of English/English retrieval using 200 cluster centers

found from clustering, removing low information words.

Name	Positives	AUC
Boxing	117	0.984
Baseball	245	0.982
Bedroom	215	0.976
Children	323	0.970
Ocean	616	0.962

Table 4.3: Top 5 AUC of English/English retrieval using seed words

Name	Positives	AUC
Inside	1591	0.824
Photograph	2463	0.813
White	4311	0.805
Picture	3487	0.721
Large	3870	0.710

Table 4.4: Bottom 5 AUC of English/English retrieval using seed words

The seed approach gives us a viable supervised approach to the keyword spotting task, giving us the flexibility of specifying which words to look for.

English-Hindi Keyword Spotting

Finally, we present the cross lingual keyword spotting results. We search for Hindi words that translate to the English seed words. We have two evaluations in the Tables 4.5 and 4.6:

1. *Hindi AUC*. This is calculated using Google Translate's English translations of the **Hindi** caption of the example found by the query word to decide whether it contains our query word.

2. *English AUC*. Each example comes with a caption pair. Here, if a Hindi example

is found by the query word, we look in the text of the **English** caption paired with the found Hindi caption to see if the query word is present.

Name	Positives	Hindi AUC	English AUC
Canyon	114	0.976	0.937
Baseball	245	0.961	0.982
Boxing	117	0.932	0.984
Bedroom	215	0.929	0.976
Snow	945	0.887	0.956

Table 4.5: Top 5 ROC of English/Hindi retrieval using seed words

Name	Positives	Hindi AUC	English AUC
Body	889	0.556	0.924
Brick	904	0.554	0.872
Large	3870	0.542	0.710
Wooden	1000	0.519	0.857
Woman	1443	0.515	0.947

Table 4.6: Bottom 5 ROC of English/Hindi retrieval using seed words

We see that the AUC calculated after performing a translation to Hindi performs worse. An intuitive explanation could be that the translation does not correspond exactly to the English word. This actually is the case. To demonstrate this, we take an example of the *building* query word.

Table 4.7: English/Hindi retrieval using *building* as a seed word

Name	Positives	Hindi AUC	English AUC
Building	3162	0.761	0.951

Taking close look at the English translations of the retrieved Hindi words, we see that there are indeed words that are semantically similar but don't correspond exactly to the seed word.

Table 4.8: Examples of Hindi words found using *building* as a seed word

बिल्डिंग : building	भवन : the building	इमारत : the building	घर : home	इमारतें : building
बंगला : bungalow	गिरजाघर : the cathedral	मकान : house	होटल : hotels	मंदि : temple

4.3 Keyword Spotting Summary

In this chapter, we presented a method for performing cross-lingual keyword spotting between English and Hindi but using a keyword from one language to find occurrences of semantically similar words in either the same language or the other language. We ran experiments in the case where the keyword was specified with ground-truth examples as well as experiments with where the keywords were derived automatically via clustering of speech segments. We used an Area Under Curve evaluation technique and showed that the method can have better performance on our selected evaluation method if we allow for synonyms.

Chapter 5

Bilingual Lexicon Discovery From Visually Grounded Speech

In this chapter, we present an approach to discover word clusters and establish their linkages from a bilingual dataset of spoken image captions. We use two different models, DAVEnet and ResDAVEnet models described in Chapter 3, and show parameter selection for both in the clustering stage. We finally present a bilingual translation table and a sample within-cluster breakdown of words found. This method proved simpler and clearer than the keyword spotting approach in Chapter 4, while providing good results. Figure 5-1 summarizes the steps used in this approach.

5.1 Method For Bilingual Lexicon Discovery

The method can be summarized as computing audio-to-audio matchmaps separately for each language and selecting regions of high similarity to obtain pseudo-words of interest, clustering the embeddings in these regions to get pseudo-word groups, and finally, linking pseudo-word groups across the two languages.

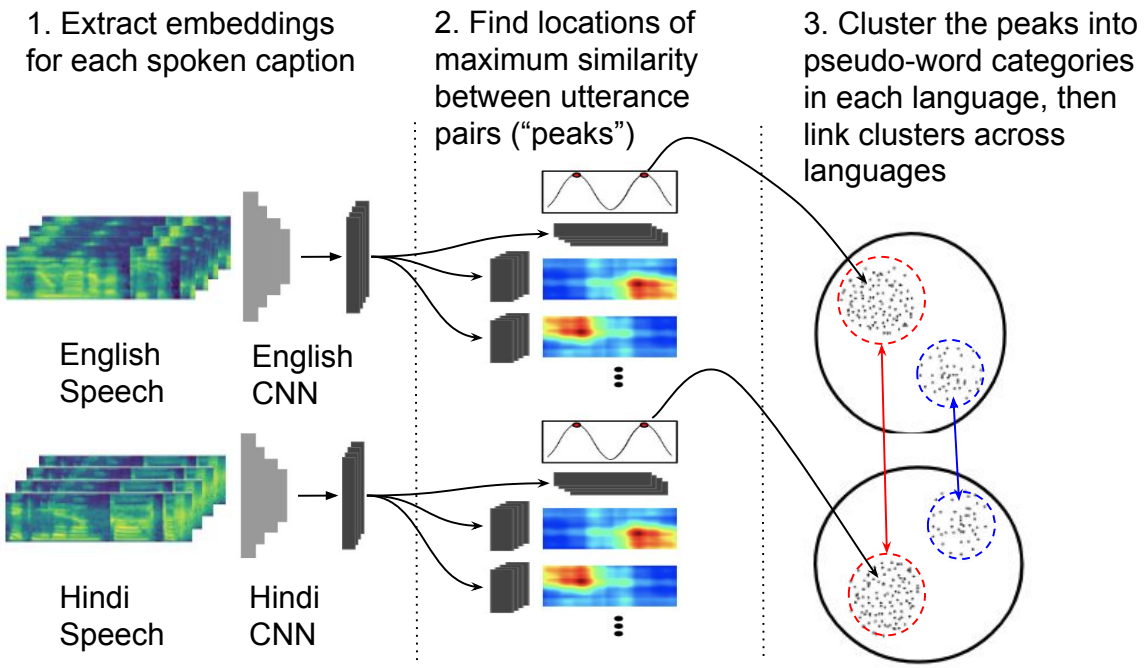


Figure 5-1: Diagram representing our proposed method for multilingual spoken lexicon discovery. Speech CNN embeddings (left) are compared to each other using dot product to find and extract regions of high similarity from utterances containing similar concepts (middle). The embeddings at these regions are clustered separately for each language and linked using cross-lingual cluster centroid similarity (right).

5.1.1 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 [39], which identifies regions of an utterance that exhibit high similarity with regions in many other utterances. While the S-DTW

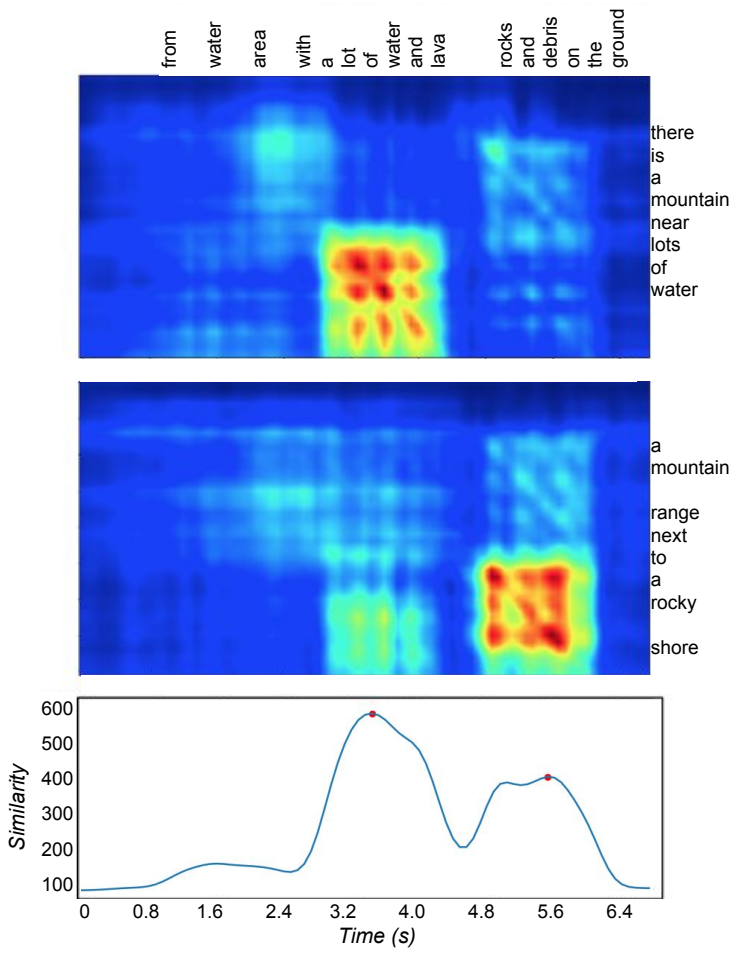


Figure 5-2: Embeddings at the speech network output for the top utterance are compared with those from its nearest neighbors. Taking the frame-level maximum over all neighbors results in the similarity profile.

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 $K = 100$.

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 5-2. To facilitate peak picking, we apply a Gaussian smoothing filter to p with $\sigma = 1$. When the word of interest occurs at the edges of the similarity profile, the profile might rise at the edge but not form a concave peak (that is, the peak only begins falling on the left side without rising first or rises but doesn't fall on the right side). We therefore add an extra point to the left and right of the profile, whose amplitude is the minimum value in the profile. To select the peaks, we use comparison among neighboring values to find local maxima. The prominence of a peak is a measure of how much it stands out due to its height and its location relative to other peaks. Given a profile, with range of values $r = max(p) - min(p)$, we select a peak if its prominence is at least $max(200, 0.15 \times r)$, since peaks with prominence less than 200 tended to not correlate with the occurrence of words of interest especially when much larger peaks exist in the profile.

5.1.2 Clustering Regions of Interest

The embedding vectors in the locations of the selected peaks now 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. Next, we perform clustering on the embedding vectors extracted at the location of each peak to get groups of major words in the two languages separately. We used a Dirichlet Process Gaussian Mixture Model clustering algorithm [8] with the following steps and parameters. Both English and Hindi peak embeddings were normalized together to have zero mean and unit variance in each dimension and then projected from $\mathbb{R}^{1024}$ to $\mathbb{R}^{300}$ using PCA. The selection of parameters for this clustering algorithm is presented in Tables 5.1 and 5.2. We cluster the two languages separately and then link them afterwards because we found that clustering them jointly resulted in separate clusters for English and Hindi speech segments.

5.1.3 Linking English and Hindi clusters

To control for repeated clusters, we selected the cluster centroids into matrices E and H for English and Hindi respectively, each with rows representing all centroid vectors in the language. We constructed an undirected graph whose edges connect every row of E with every row of H and the edge weights represented by the dot product of E_i and H_j. We then used a threshold τ below which edge weights are set to zero. The selection of the value of τ is presented in Table 5.3. Finally, we ran a graph clustering algorithm to group the DPGMM English and Hindi cluster centroids into meta-clusters that capture bilingual concepts. We used the Louvain community clustering [9] for this step.

5.1.4 Deriving cluster labels for evaluation

Each peak at the $conv5$ layer has a receptive field of size f frames with respect to the spectrogram input. Since the neural network exhibits symmetry in both convolution and pooling operations, we simply find the location in the speech caption directly

below the peak as $p \times c/n$, where p is the location of the peak in the last convolution layer, n is the number of frames in the last convolution layer and c is the caption length in seconds. We then snip $f/2$ frames (corresponding to s 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 thesis. While this method of deriving peak labels is bound to capture some spurious words that did not trigger the activation of the peaks, it is simple and effective. 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 frequent 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 for DAVEnet audio channel and 8.4 seconds for ResDAVEnet audio channel. We used 2.5 seconds in all our experiments since in practice, it provided good results for both networks.

5.2 Experiments

In this section, we present the hyperparameters selection process for Dirichlet Process Gaussian Mixture Model, used in discovering word clusters for both DAVEnet and ResDAVEnet. Then we present the edge threshold selection used for finding meta clusters for ResDAVEnet. Finally, we show a quantitative performance of the bilingual lexicon discovery method as well as a sample breakdown of some word clusters.

5.2.1 Setup

We first select parameters to use in the clustering process. Afterwards, we select parameters for meta-clustering in this section. In general, we judged good clusters by their F_1 score, defined as:

$$F_1 = 2\frac{purity \cdot coverage}{purity + coverage}$$

The closer the F_1 score is to 1, the better the model is at reporting pure pseudo-word clusters while capturing more of the occurrences of the relevant words in the dataset.

DPGMM Selection for DAVEnet

The first rows of Tables 5.1 and 5.2 show the default parameter settings. All other rows show the parameters that were changed for each experiment. *CVT* refers to the type of covariance matrix used for each component, *TOL* is the change in average likelihood of the data with respect to the model below which we take the training to have converged, *MPP* defines the extend means can be placed (larger values concentrate the cluster means around the mean priors), N is the number of seed clusters, *WCP* is the Dirichlet concentration of each component on the weight distribution (lower values lead to more sparse solutions for active components), *MI* is the maximum number of iterations to run for the Expectation Maximization algorithm.

DPGMM Selection for ResDAVEnet

We found both models to have the same parameter selection for DPGMM clustering. In general, ResDAVEnet, which had higher recall scores of the two models also found more clusters with F_1 scores above 0.5, indicating its ability better find quality word clusters.

Selection for Meta-Cluster Edge Weight Pruning Threshold

After using the best selection of parameters for DPGMM clustering, we proceed to group the clusters into meta clusters in order to group similar clusters across languages as well as reduce the repetition of concepts. We find the best value of the edge weight pruning to use in the graph clustering in order to maximize the clusters with both

CVT	TOL	MPP	N	WCP	MI	CLUSTERS	>.5	>.4	>.3
diag	0.001	1	200	1	1500	146	57	89	114
				1000		146	57	89	114
				100000		146	57	89	114
		0.1				157	53	84	109
		10				151	61	94	120
		50				**161**	**64**	**101**	**126**
		100				160	63	99	123
		150				154	58	96	119
		200				152	57	95	116
		1000				111	46	69	84
			1000			145	53	79	103
		50	1000			182	63	99	127
		100	1000			181	62	100	126
	1.00E-05					146	57	89	114
	1.00E-10					146	57	89	114

Table 5.1: Sensitivity of Dirichlet Process Gaussian Mixture Model parameters using DAVEnet model. The legend for the column headers is *Covariance Type (CVT)*, *Tolerance (TOL)*, *Mean Precision Prior (MPP)*, *Number of seed clusters (N)*, *Weight Concentration Prior (WCP)*, *Maximum Iterations (MI)*, *Number of clusters found (CLUSTERS)*. The last three columns refer to the number of clusters with F_1 Score above 0.5, 0.4 and 0.3 respectively.

CVT	TOL	MPP	N	WCP	MI	CLUSTERS	>.5	>.4	>.3
diag	0.001	1	200	1	1500	98	68	79	88
				1000		95	68	79	88
				100000		95	68	79	88
		0.1				101	67	78	87
		10				110	77	91	103
		50				**127**	**88**	**106**	**121**
		100				127	87	104	119
		150				126	85	103	117
		200				121	79	98	113
		1000				89	59	74	87
			1000			84	60	71	77
		50	1000			123	86	101	113
		100	1000			126	86	98	113
	1.00E-05					95	68	79	88
	1.00E-10					95	68	79	88

Table 5.2: Sensitivity of Dirichlet Process Gaussian Mixture Model parameters using the ResDAVEnet model. The legend for the column headers is *Covariance Type (CVT)*, *Tolerance (TOL)*, *Mean Precision Prior (MPP)*, *Number of seed clusters (N)*, *Weight Concentration Prior (WCP)*, *Maximum Iterations (MI)*, *Number of clusters found (CLUSTERS)*. The last three columns refer to the number of clusters with F_1 Score above 0.5, 0.4 and 0.3 respectively.

English and Hindi cluster centroids (pseudo-translations).

English			Hindi								τ
>.5	>.4	>.3	>.5	>.4	>.3	N_E	N_H	NU_E	NU_H	N_T	
0	10	16	23	10	19	24	30	29	30	27	29
100	8	16	28	12	18	24	33	33	33	29	32
200	11	22	33	15	28	34	42	41	41	39	39
300	13	27	40	20	32	39	47	48	46	45	41
400	**24**	**38**	**53**	**30**	**46**	**54**	**62**	**65**	**65**	**65**	**53**
500	29	42	57	31	51	58	68	68	66	62	47
600	38	57	72	37	55	67	81	78	79	70	45
700	55	73	88	39	58	70	98	81	96	73	32
800	66	86	100	39	59	71	111	82	107	74	28
900	75	93	108	39	59	71	119	82	115	74	24
1000	77	96	111	39	60	71	122	83	117	75	20
1100	79	98	113	39	60	71	124	83	119	75	16
1200	81	100	115	41	61	72	126	84	121	76	11

Table 5.3: Sensitivity of meta clustering to edge weight thresholding. N_E is the number of meta clusters with English words, N_H is the number of meta clusters with Hindi words, NU_E and NU_H refer to the number of unique cluster names in English and Hindi respectively. N_T is the number of translations (meta clusters with both English and Hindi). >.5, >.4 and >.3 refer to the number of clusters within the language with F_1 Score above 0.5, 0.4 and 0.3 respectively using the best performing DPGMM result.

5.2.2 Results

Bilingual Speech Lexicon Discovery

To show the bilingual lexicon discovery we follow the methodology in Section 5.1 to find clusters in the two languages independently and then link them afterwards. We present the top ten word level translations in Table 5.4. Appendix A.1 contains a more complete list of the discovered English-Hindi pseudo-word pairs. 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 translations 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 separately. In our table, we merged the statistics of singular and plural versions of words in the clusters. It is clear from the table that we successfully link words from the two languages automatically.

Table 5.4: 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.

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
boat(s)	0.64	sailboat	0.04	नाव:the boat	0.24	'जहाज':ship	0.04	1694	1317	1229	0.60	0.76
bridge	0.77	suspension	0.04	पुल:the bridge	0.27	ब्रिज:the bridge	0.23	1805	1681	895	0.64	0.21
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

Word Selection

The clusters found by our method groups together semantically related words. As shown in Table 5.5 our current method finds variations of the same word or in some cases adjectives and verbs that go with the top cluster label. Since we have a wide receptive field for getting clusters, we capture adjectives and verbs which would not otherwise necessarily belong to the cluster. *Mountains* for instance within the *snow cluster* is most like because of phrases like *snowy mountain*. Using a finer-grained segment extraction method like DTW-based alignment would reduce the number of spurious words included in the clusters.

Bilingual Picture Dictionary

We show a few examples of the bilingual picture dictionary in Figure 5-3. The picture dictionary is created using the VGG model from the DAVEnet training procedure. It is easy to note that the model focuses on portions of the image corresponding to the words in both English and Hindi.

Snow cluster			Bedroom cluster			Boat Cluster		
Word	N	Purity	Word	N	Purity	Word	N	Purity
snow	1758	0.66	bedroom	416	0.34	boat	744	0.53
snowy	460	0.17	bed	589	0.47	boats	285	0.20
mountain	368	0.14	beds	125	0.1	water	220	0.16
mountains	209	0.08	room	140	0.11	sailboat	45	0.03
snowing	32	0.01	inside	48	0.04	harbor	25	0.02
snowman	20	0.008	pillows	29	0.02	ocean	27	0.02
igloo	21	0.008	walls	21	0.02	sailboats	16	0.01
skiing	16	0.006	blanket	16	0.01	marina	20	0.01
snowboard	10	0.002	sheets	14	0.01	docked	20	0.01
snowboarder	8	0.003	wooden	12	0.01	lake	17	0.01

Table 5.5: Concepts found within clusters presented for three clusters. N refers to the number of instances of the word present anywhere in speech segments found in the given cluster.

5.3 Bilingual Lexicon Discovery Summary

We introduced an interval piling approach to independently and automatically find word groups separately in two languages and link them, resulting in a bilingual picture dictionary. In addition, we showed that model recall generally correlates positively to the number of unique and quality word groups discovered. Finally, we showed that our optimal parameter selection for clustering worked well for at least two different settings (two different models).

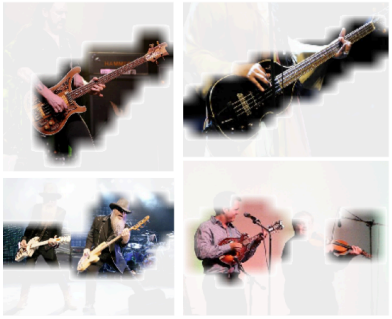
Figure 5-3: Picture dictionary representing three-way agreement between English speech caption, Hindi speech caption and Image pixels using DAVEnet. We present the text transcriptions of the clustered speech segments with their corresponding cluster purities.

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)

Chapter 6

Trilingual Lexicon Discovery

With the current setup for bilingual lexicon discovery, one follow up question is whether it is possible to have agreement between more than two languages with minimal changes. Further, we wanted to know whether adding a grounding from a third language could improve word discovery. In this chapter, we introduce the Japanese caption dataset for the Places images followed by a brief description of statistics about the three language captions. We then present modifications to our model and meta clustering algorithm in order to accommodate the third language. Finally, we present results from experiments comparing monolingual, bilingual and trilingual cases for lexicon discovery.

6.1 Dataset Extension

We use a Japanese dataset on a similar subset of places image as the Hindi description. The captions are obtained from [1]. In total, after selecting all the images shared among the three languages, we had 74,000 images with captions in all three languages for training and another 1K for validation. The validation set from the Bilingual case had only 893 images in common with the Japanese dataset. We therefore added an additional 107 images common to all three language captions to keep the validation set at 1K. This means that the validation set used here is different from that used in Chapter 5. However, for this chapter, the validation set of images is fixed across

monolingual, bilingual and trilingual cases. In order to present our results in text form, we use Google's Automatic Speech Recognition API to transcribe the Japanese captions and use the Google Translate API to get the English translations of the Japanese captions.

6.1.1 Data Statistics

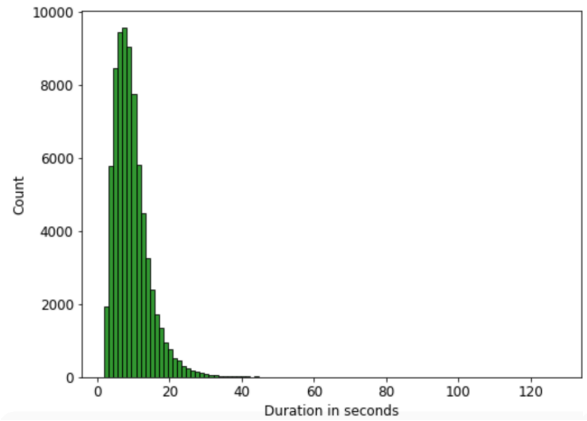
Language	Mean	Standard deviation	min	max
English	9.58	5.11	2.01	127.76
Hindi	7.54	4.42	0.04	110.29
Japanese	19.74	9.17	4.10	204.54

Table 6.1: Statistics of audio caption lengths in seconds for each language.

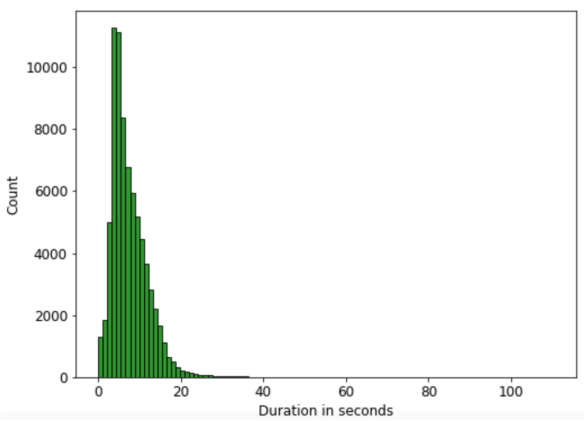
Japanese captions in our dataset were much longer than English and Hindi and had more spread, as shown in Table 6.1 and the accompanying caption length histogram is presented in Figure 6-1. In addition, from the ASR transcripts of the caption, we provide a frequency count of words in each language. The Japanese dataset had an average of 33.61 words per caption, whereas English and Hindi had 20.50 and 20.75 respectively. The complete distribution of number of words in different frequency bins is shown in Figure 6-2. We therefore expect to find more Japanese words during the lexicon discovery process compared to the other two languages.

6.2 Model Extension

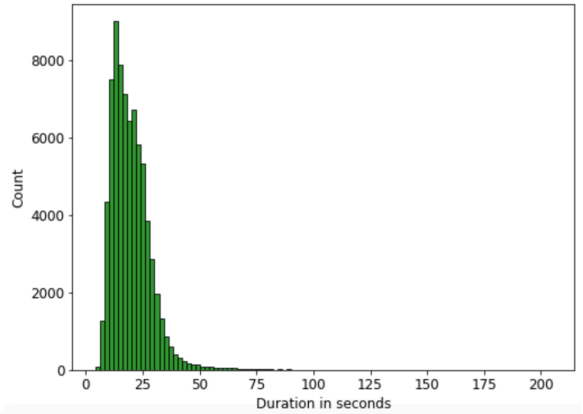
To accommodate for a third language caption input to the network, we simply add a third DAVEnet model, totalling 4 CNN models. The loss function is updated as well to reflect the new language. In all, there are 12 terms in the new loss function, representing all possible combination of image and audio retrieval loss. Specifically, we have E↔I, E↔H, E↔J, H↔I, H↔J, J↔I, where E, H and J represent English, Hindi and Japanese respectively and I represents the image. All audio-audio losses are weighted by a factor of 5 in the loss function whereas audio-image losses are weighted



(a) English caption durations



(b) Hindi caption durations



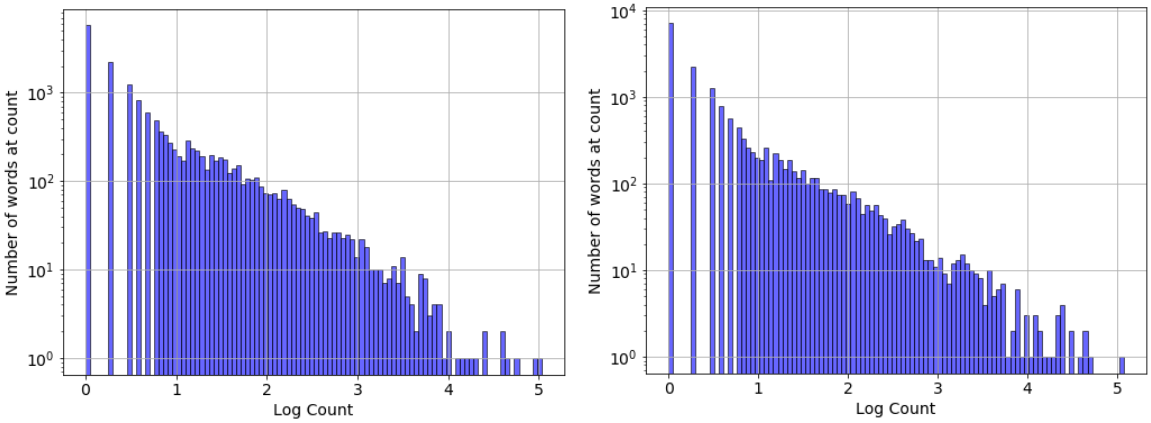
(c) Japanese caption durations

Figure 6-1: Histograms of caption durations by language. Japanese captions were on average longer than those of English and Hindi, with the Japanese histogram having a longer tail.

by a factor of 1. Just as in the biligual case, all models were trained in one round of 90 epochs, with batch size of 128 and initial learning rate of 0.001, decreasing by a factor of 10 every 25 epochs. We decrease the learning rate every 25 instead of 30 epochs to accommodate for training in the trilingual setting which tended to diverge when the initial learning rate was sustained for longer.

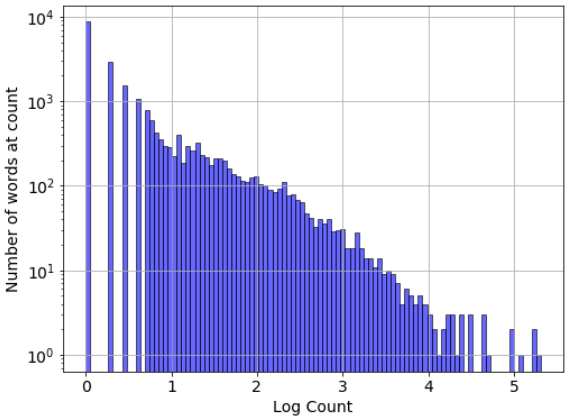
6.3 Clustering Extension

Just as in the bilingual case, we process each language separately at the matchmap creation step. We then project the found peaks in all the languages from R^{1024} to R^{300}



(a) English words statistics

(b) Hindi words statistics



(c) Japanese words statistics

Figure 6-2: Number of words (y axis) which occur log number of times (x axis) in the dataset by language. Japanese captions had more unique words, which occurred more frequently than those of English and Hindi, and had more very high frequency words as well.

using the same principal components (we find the PCA axes using data points from all three languages at the same time). Then we perform clustering on the languages separately. After we have the within language pseudo-word groups, we make an extension to the meta clustering process for three languages. We create a matrix, M whose rows are the cluster centroids for all the languages, stacked into a single matrix. That is, the first N_E rows contain all the English cluster centroids, the next N_H rows contain the Hindi centroids, and the last N_J, the Japanese cluster centroids, where N_E, N_H, N_J represent the number of clusters found for English, Hindi and Japanese respectively. From this matrix, we make an adjacency matrix $A = MM^T$ and the

entries represent the dot product similarity between each cluster centroid and all other cluster centroids. We set edge weights below 400 to zero in the trilingual case. Next, we perform graph clustering using the Louvain community clustering algorithm [9] and report meta-clusters that capture centroids from all three languages.

Let a meta-cluster capture centroids $\{C_{E_1}, C_{E_2}, \cdots, C_{N_E}, C_{H_1}, C_{H_2}, \cdots, C_{N_H}\}$ where N_E and N_H refer to the number of English and Hindi centroids respectively in the meta cluster. In the bilingual case, we defined cross-lingual similarity between English and Hindi as:

$$S_{EH} = \langle \frac{1}{N_E} \sum_{i=1}^{N_E} C_{E_i}, \frac{1}{N_H} \sum_{j=1}^{N_H} C_{H_j} \rangle$$

where $\langle \cdot, \cdot \rangle$represents the inner product operation. In the trilingual case, we extend this definition by finding the average of S_{EH}, S_{EJ} and S_{HJ}, assuming that the meta-cluster contains $\{C_{J_1}, C_{J_2}, \cdots, C_{N_J}\}$. J refers to Japanese and N_J is the number of Japanese clusters in the meta-cluster.

6.4 Experiments

In this section, we present the change in performance of the word discovery model in the monolingual, bilingual and trilingual settings. We first show the recall scores obtained by the different models and then the number of words found by the different models. As in the bilingual case, we also show the trilingual lexicon discovered in Table 6.7. We finally present a picture dictionary in Appendix B-1

6.4.1 Setup

For our experiments, we initialized all audio channels of the neural network with weights from a ResDAVEnet model trained on 400K English captions. The actual training process used in this chapter used the 74K intersection between all three languages. Since the English corpus is much larger than both Hindi and Japanese and the English dataset was used in pre-training the speech models, we make the

language level comparisons on Hindi and Japanese datasets. All experiments in this section use Resnet50 and ResDAVEnet.

i. We train a monolingual model for each of Hindi and Japanese using English 400K pretrained model.

ii. We train bilingual models for Hindi-English and Japanese-English

iii. We train a trilingual model for English-Hindi-Japanese

We use the same clustering algorithm across all experiments. To allow for fair comparison, we present changes in the number clusters found with F1 score greater than (.5, .4, .3) for Hindi and Japanese in the monolingual, bilingual and trilingual settings. We then present a separate analysis for English on whether the addition of the different languages increased the number of words with high F1 scores within the same 75K subset of training data.

6.4.2 Comparing Monolingual, Bilingual and Trilingual Cases

We compare the number of words found by the Hindi model and the Japanese models when trained separately versus trained jointly. Both models were initialized using a pretrained English audio model, trained on 400K image audio pairs. As shown in Table 6.2, the recalls generally increased for each language with increasing number of languages. Hindi had the lowest performance of the three languages in most cases. This is possibly due to the fact that the Hindi captions were shorter as we describe in Section 6.4.2.

As seen in Table 6.2, switching from monolingual to bilingual increased the recalls for both Hindi and Japanese. Although the trilingual setting increases the recalls of both languages, the gain is marginal compared to the bilingual case. The Hindi-Japanese recall however gets a larger boost in the trilingual case compared to training Hindi- Japanese bilingual. Compared to adding English to each of Hindi and Japanese in a bilingual setting, using the Hindi-Japanese pairing in training led to lower final recalls. This could be because the English channel is easier to train given that it was

Model	Recalls@10	Recalls@5	Recalls@1
Monolingual Hindi			
I→H	0.446	0.324	0.133
H→I	0.450	0.333	0.118
Monolingual Japanese			
I→J	0.583	0.442	0.176
J→I	0.603	0.479	0.213
Bilingual Hindi-English			
I→H	0.590	0.468	0.189
H→I	0.594	0.490	0.201
Bilingual Japanese-English			
I→J	0.724	0.581	0.238
J→I	0.757	0.626	0.302
Bilingual Hindi-Japanese			
I→H	0.570	0.458	0.171
H→I	0.584	0.474	0.212
I→J	0.676	0.520	0.191
J→I	0.720	0.597	0.266
J→H	0.422	0.323	0.108
H→J	0.423	0.321	0.141
Trilingual Hindi-Japanese-English			
I→H	0.599	0.474	0.201
H→I	0.606	0.477	0.222
I→J	0.725	0.576	0.234
J→I	0.781	0.654	0.291
J→H	0.475	0.353	0.118
H→J	0.493	0.373	0.133

Table 6.2: Comparing retrieval scores for Hindi and Japanese in the monolingual, bilingual and trilingual settings. I, H and J refer to Image, Hindi and Japanese respectively and the right arrow shows that the item on the left is used to retrieve the item on the right.

pretrained with English. Hyperparameter tuning (of weights in the loss function) gets more complex as the more languages are added, which could be one reason why performance went up for some languages but down for others.

Table 6.3 suggests that increasing recalls of the model generally increases the number of quality clusters found. Clusters with F1 scores greater than 0.5 are purer and have good coverage over the dataset, therefore representing underlying concept

well. As expected, Hindi had lower monolingual recall scores compared with Japanese and therefore found fewer high quality clusters than Japanese. Japanese performed well even when the model was initialized with weights trained on English captions in the monolingual case, possibly due to longer Japanese captions per training example. As suggested from the recalls presented, the number of words found generally increases more substantially from monolingual to bilingual case, and improves marginally in the trilingual setting.

Language	Setting	Number of clusters	F1>.5	F1>.4	F1>.3
Hindi	H	18	4	8	10
Hindi	H-E	84	43	60	67
Hindi	H-E-J	88	44	66	70
Japanese	J	49	21	28	41
Japanese	J-E	151	88	118	137
Japanese	J-E-H	154	85	112	129

Table 6.3: Comparing Number of words found for Hindi and Japanese in the monolingual, bilingual and trilingual settings as well as number of clusters with F1 scores greater than three different thresholds.

Since the English branch of the network benefits from pretraining from a substantially larger dataset than the other two languages (400K vs 74K), we present the English statistics separately with recalls in Table 6.4 and word clusters in Table 6.5.

Effect of caption length constraint on recalls

To verify that the Japanese captions perform well in the retrieval task largely due to the longer caption lengths, we trained a trilingual model while constraining the input captions to the first 7.5 seconds (the smallest average caption length of the three languages). This constraint was also imposed on the validation set. After training using a similar setup as our previous trilingual experiments, we achieved the recalls shown Table 6.6. It is clear that no language performs disproportionately better than the others in the retrieval task, beside English which was used in pretraining all the audio networks used in the experiments.

Model	Recalls@10	Recalls@5	Recalls@1
Monolingual English			
I→E	0.704	0.575	0.260
E→I	0.740	0.620	0.306
Bilingual English-Hindi			
I→E	0.658	0.522	0.209
E→I	0.684	0.558	0.260
H→E	0.419	0.318	0.128
E→H	0.422	0.326	0.129
Bilingual English-Japanese			
I→E	0.663	0.520	0.218
E→I	0.682	0.555	0.267
J→E	0.477	0.366	0.128
E→J	0.490	0.382	0.128
Trilingual English-Hindi-Japanese			
I→E	0.695	0.545	0.224
E→I	0.700	0.587	0.291
H→E	0.439	0.343	0.148
E→H	0.435	0.327	0.142
J→E	0.502	0.360	0.129
E→J	0.497	0.379	0.141

Table 6.4: Comparing retrieval scores for English in the monolingual, bilingual and trilingual settings.

Language	Setting	Number of clusters	F1>.5	F1>.4	F1>.3
English	E	93	48	65	75
English	E-H	127	88	106	121
English	E-J	137	98	117	132
English	E-H-J	134	92	111	124

Table 6.5: Comparing Number of words found for English in the monolingual, bilingual and trilingual settings as well as number of clusters with F1 scores greater than three different thresholds.

6.4.3 Cross-Lingual Word Linkage

We present the top meta clusters by within cluster similarity for trilingual word level translation in Table 6.7. Please refer to Appendix A.2 for a more extensive list of

Model	Recalls@10	Recalls@5	Recalls@1
Trilingual English-Hindi-Japanese			
I→E	0.608	0.472	0.189
E→I	0.610	0.489	0.214
I→H	0.555	0.423	0.156
H→I	0.551	0.421	0.174
H→E	0.382	0.296	0.117
E→H	0.404	0.299	0.102
I→J	0.565	0.438	0.172
J→I	0.568	0.451	0.182
E→J	0.369	0.279	0.099
J→E	0.356	0.278	0.109
J→H	0.324	0.241	0.086
H→J	0.324	0.239	0.095

Table 6.6: Recalls in the trilingual setting with length constrained input captions.

word linkages.

Of the words learned in the trilingual setting, we select the top cluster labels for each language and present statistics on the number of occurrences in the dataset. We first found the frequency count of the top labels in the dataset. We then found the mean and standard deviation of the frequency counts and removed words that had frequency count greater than 2 times the standard deviation (mostly highly frequent stop words). Table 6.8 shows statistics about the number of occurrences of discovered words.

6.4.4 Picture Dictionary

We present a few of the image-caption clusters in the picture dictionary in Figure 6-3. The picture dictionary is made using the Resnet50 model from ResDAVEnet. More examples are presented in Appendix B-1

Table 6.7: Trilingual Lexicon Discovery. This table presents the top 3 cluster labels for some of the discovered trilingual clusters. p1, p2 and p3 refer to the purity scores using the top 3 labels respectively and N is the number of speech segments in each cluster.

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	**Similarity**
kitchen	0.89	a	0.40	the	0.25	0.79	1098	
रसोईघर:kitchen	0.23	रसोई:kitchen	0.25	है:is	0.34	0.88	886	2158
キッチン:kitchen	0.73	の:of	0.35	ダイニング:dining	0.03	0.70	1396	
bus	0.70	inside	0.16	a	0.48	0.57	353	
बस:bus	0.80	अंदर:inside	0.32	एक:one	0.42	0.31	250	1945
バス:bus	0.77	の:of	0.70	車内:Inside the car	0.21	0.53	374	
windmill	0.38	windmills	0.21	wind	0.21	0.83	606	
पवन:air	0.40	दिखाई:visible	0.22	चक्की:mill	0.20	0.75	464	1697
風車:Windmill	0.48	風力:Wind power	0.22	の:of	0.34	0.70	819	
boxing	0.59	ring	0.29	boxers	0.13	0.75	434	
बॉक्सिंग:boxing	0.71	रिंग:ring	0.26	एक:one	0.23	0.66	263	1566
ボクシング:boxing	0.54	リング:ring	0.30	の:of	0.48	0.89	491	
black	0.85	white	0.82	photo	0.32	0.22	1383	
तस्वीर:picture	0.48	ब्लैक:the black	0.32	यह:this	0.26	0.05	496	1542
白黒:Black and white	0.62	モノクロ:Monochrome	0.22	写真:Photo	0.50	0.91	2065	
horse	0.50	horses	0.30	a	0.39	0.70	627	
घोड़े:the horse	0.32	घोड़ा:horse	0.12	है:is	0.21	0.76	696	1524
馬:Horse	0.55	競馬:Horse racing	0.13	の:of	0.33	0.49	707	
bicycle	0.20	motorcycle	0.17	bike	0.14	0.77	459	
साइकिल:bicycle	0.34	मोटरसाइकिल:motorcycle	0.09	है:is	0.23	0.78	603	1512
マイク:Microphone	0.20	自転:rotation	0.20	バイク:bike	0.17	0.66	1129	
bedroom	0.30	bed	0.41	a	0.42	0.85	1241	
बिस्तर:bed	0.62	है:is	0.34	सोने:the gold	0.21	0.73	847	1287
ベッド:bed	0.63	の:of	0.35	が:But	0.30	0.51	1917	
statue	0.44	statues	0.17	sculpture	0.11	0.75	902	
मूर्ति:cculpture	0.47	मूर्तियां:sculptures	0.24	है:is	0.34	0.69	529	1247
銅像:Bronze statue	0.12	の:of	0.45	が:But	0.33	0.76	1789	
fish	0.54	station	0.26	gas	0.28	0.65	731	
मछली:fish	0.47	मछलियां:fish	0.23	है:is	0.21	0.63	469	1212
ガソリン:gasoline	0.23	魚:fish	0.29	水槽:Water tank	0.13	0.63	1317	
staircase	0.25	stairs	0.26	steps	0.15	0.84	1135	
सीढ़ियां:stairs	0.27	चिड़िया:bird	0.14	है:is	0.27	0.93	1342	1134
階段:Stairs	0.61	の:of	0.33	が:But	0.24	0.64	2285	
asian	0.43	chinese	0.21	building	0.17	0.39	588	
मंदिर:temple	0.47	है:is	0.24	एक:one	0.21	0.38	453	1107
五重塔:Five-story pagoda	0.09	の:of	0.56	寺:temple	0.18	0.83	710	
boat	0.51	boats	0.20	a	0.39	0.66	1396	
नाव:the boat	0.28	है:is	0.29	जहाज:ship	0.16	0.76	1225	1092
船:boat	0.36	ボート:boat	0.20	の:of	0.35	0.58	2175	
baseball	0.42	course	0.21	golf	0.27	0.76	1481	
खिलाड़ी:the player	0.42	मैदान:field	0.19	यहां:here	0.11	0.75	1107	997
ゴルフ:golf	0.30	野球:baseball	0.28	の:of	0.44	0.68	1966	
children	0.26	child	0.12	young	0.17	0.76	2088	
बच्चे:children	0.43	बच्चा:child	0.16	छोटे:small	0.17	0.58	2003	926
子供:children	0.32	男の子:boy	0.13	女の子:girl	0.14	0.62	3287	
ocean	0.53	beach	0.35	the	0.57	0.79	1661	
समुद्र:sea	0.28	समंदर:ocean	0.25	किनारे:edge	0.26	0.66	2882	876
海:Sea	0.48	砂浜:Sandy beach	0.08	海岸:Coast	0.06	0.64	4676	

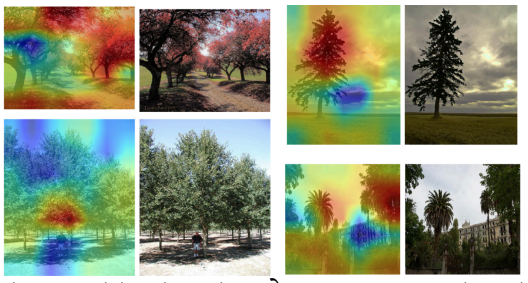
Language	Mean	Standard deviation	min	max
English	2118	2218	128 (bicycle)	9634 (white)
Hindi	3239	6707	228 (बॉक्सिंग, boxing)	49303 (एक, one)
Japanese	2908	3407	78 (五重塔, five-story pagoda)	20087 (写真, photo)

Table 6.8: Statistics of number of occurrences of words for the top labels in the dataset for each language

6.5 Trilingual Lexicon Discovery Summary

In this chapter, we were able to show that the bilingual lexicon discovery method proposed in Chapter 5 extends to more than two languages with only minimal changes. We presented a comparison of the model performances in the monolingual, bilingual and trilingual cases as well as examined the effect of caption length on the discovery capability. Finally, we presented a trilingual translation table as well as a picture

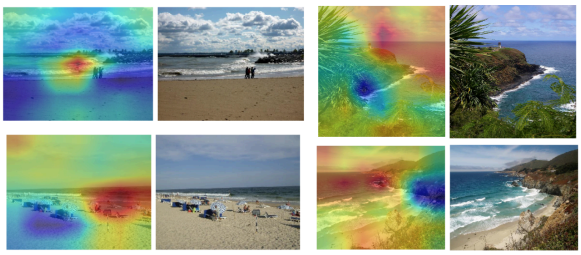
Figure 6-3: Picture dictionary representing four-way agreement between English speech caption, Hindi speech caption, Japanese speech caption and Image pixels using ResDAVEnet. We present the text transcriptions of the clustered speech segments with their corresponding cluster purities.



(a) tree(s) (0.79), पेड़:the trees (0.59), 木:wood (0.46)

(b) horse(s) (0.80), घोड़े:the horse (0.44), 馬:Horse (0.68)

(c) car(s) (0.53), गाड़ी:the cart (0.71), 車:car(0.43)



(d) ocean (0.53), समुद्र:sea (0.53), 海:Sea (0.48)

dictionary.

Chapter 7

Conclusion

7.1 Summary of Contributions

In this thesis, we presented two approaches for discovering word-like units in visually-grounded speech. After applying our method to the monolingual case, we showed that semantically similar words across two languages can be automatically linked using the embedding space created by the pre-global pooling layer of the DAVEnet model. We then extended the work to the triligual case to confirm that linking three languages together creates a three-way word level translation among the three languages. Our initial work only showed significant improvements in the number of words found from the monolingual to bilingual setting and marginal improvement in the trilingual case, although further experiments might reveal added benefits of using a third language. Finally, we presented tables of the learned words in both the bilingual and trilingual settings as well as regions of images they describe, making a picture dictionary.

7.2 Future Work

Our current approach requires us to utilize spoken captions for a common set of images for both languages, but in the future we plan to investigate whether similar results can be achieved when different sets of images are used for each language's captions. Future work should also investigate direct speech-to-speech translation using

our discovered meta-clusters. We would also like to refine the segmentation of the speech captions as the fixed 2.5 seconds window is a potential source of errors. In addition, expanding the number of words discovered past the salient objects, adjectives and verbs will substantially increase the promise of this method. Finally, we believe that the representations learned by our acoustic models could find use in traditional speech recognition systems, such as in low-resource or code switching scenarios.

7.3 Parting Thoughts

As a speaker of a low resource language, I am thrilled to have been part of a new direction in spoken language processing research. I look forward to seeing the techniques explored in this paper further improved for larger scale building of linguistic units at a much finer granularity which can further be used in natural translation between multiple languages. Furthermore, I hope future work can work to reduce major constrains and heuristics used in training language models and segmenting and classifying linguistic units, such as the need for aligned captions from multiple languages for each image. Such a breakthrough will open many more possibilities, especially in this age where large quantities of videos in several languages are easily accessible on the internet. Maybe one day, I will see translations from Ghanaian dialects to other languages without the need for expert linguist supervision.

Appendix A

Tables

Table A.1: Bilingual word clusters with each block representing English and Hindi clusters within a meta-cluster. $p1$, $p2$ and $p3$ refer to the purity scores of the top three words in the cluster. N is the number of peaks in the meta-cluster for each language.

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
lighthouse	0.55	house	0.28	white	0.25	0.73	485	
लाइट:light	0.36	हाउस:house	0.30	है:is	0.28	0.30	578	4898
black	0.62	white	0.55	old	0.21	0.23	2114	
तस्वीर:picture	0.21	पुरानी:old	0.18	एक:one	0.26	0.06	1479	3340
bedroom	0.29	bed	0.33	a	0.42	0.85	1620	
बिस्तर:bed	0.37	है:is	0.26	सोने:the gold	0.19	0.73	1535	2470
guitar	0.71	playing	0.39	guitars	0.08	0.74	280	
गिटार:guitars	0.55	है:is	0.29	बजा:time	0.22	0.76	411	1961
staircase	0.21	stairs	0.23	steps	0.16	0.60	1115	
सीढ़ियां:stairs	0.25	चिड़िया:bird	0.14	है:is	0.28	0.81	1305	1925
windmill	0.35	windmills	0.20	wind	0.20	0.69	569	
पवन:air	0.47	चक्की:mill	0.23	चक्कियां:chakkiy	0.12	0.74	627	1882
kitchen	0.82	a	0.36	with	0.17	0.57	933	
रसोईघर:kitchen	0.17	रसोई:kitchen	0.20	है:is	0.28	0.82	1292	1827
bridge	0.77	a	0.42	suspension	0.04	0.64	1681	
है:is	0.33	ब्रिज:the bridge	0.23	एक:one	0.37	0.01	895	1805
couch	0.21	room	0.30	living	0.30	0.80	1693	
बैठने:sit down	0.29	कमरे:the room	0.26	लिए:for	0.34	0.20	895	1796
mountain	0.33	mountains	0.24	a	0.30	0.49	3009	
है:is	0.33	पहाड़:the mountain	0.21	चट्टान:rock	0.14	0.01	1920	1772

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
pool	0.55	swimming	0.10	a	0.36	0.61	1291	
तालाब:pond	0.39	है:is	0.37	स्विमिंग:swimming	0.14	0.34	864	1743
boat	0.44	boats	0.20	water	0.08	0.50	1317	
नाव:the boat	0.24	है:is	0.28	एक:one	0.23	0.62	1229	1694
flowers	0.53	garden	0.22	flower	0.09	0.72	1987	
बगीचा:garden	0.22	फूल:flower	0.30	बगीचे:the garden	0.15	0.54	1800	1642
fish	0.64	aquarium	0.07	a	0.28	0.63	639	
मछली:fish	0.49	मछलियां:fish	0.22	है:is	0.28	0.57	467	1617
stage	0.71	on	0.44	a	0.31	0.58	437	
है:is	0.54	रही:doing	0.30	और:and	0.18	0.01	1495	1589
snow	0.49	snowy	0.13	covered	0.15	0.70	3464	
बर्फ:ice	0.47	है:is	0.37	बर्फीले:snowy	0.11	0.55	3221	1586
closet	0.62	a	0.38	cabinets	0.05	0.63	406	
अलमारी:cupboard	0.51	है:is	0.47	अलमारियां:shelves	0.08	0.49	752	1537
windows	0.33	window	0.20	restaurant	0.07	0.50	2117	
दुकान:shop	0.82	है:is	0.35	एक:one	0.41	0.39	1123	1499
baseball	0.35	stadium	0.19	football	0.17	0.58	1330	
मैदान:field	0.34	कुर्सियां:chairs	0.21	है:is	0.27	0.28	3369	1469
boxing	0.36	boxers	0.09	boxes	0.09	0.78	1467	
बॉक्सिंग:boxing	0.29	मुक्केबाजी:boxing	0.11	प्रतियोगिता:contest	0.12	0.54	696	1453
horse	0.47	horses	0.27	a	0.40	0.62	674	
घोड़े:the horse	0.41	घोड़ा:horse	0.20	है:is	0.22	0.59	442	1431
bowling	0.61	alley	0.35	ball	0.15	0.63	530	
है:is	0.26	एक:one	0.19	यहां:here	0.07	0.00	603	1428
track	0.37	running	0.37	race	0.22	0.25	407	
दौड़:the race	0.31	प्रतियोगिता:contest	0.13	दौड़ते:running	0.19	0.43	408	1422
car	0.44	cars	0.14	a	0.37	0.44	1273	
गाड़ी:the cart	0.56	एक:one	0.29	है:is	0.22	0.21	943	1422
golf	0.78	course	0.46	a	0.53	0.67	692	
मैदान:field	0.22	हरियाली:greenery	0.16	है:is	0.21	0.05	621	1359
bus	0.55	inside	0.15	a	0.43	0.61	851	
बस:bus	0.40	हवाई:airy	0.24	एक:one	0.33	0.31	930	1319
train	0.71	a	0.40	trains	0.05	0.34	984	
गाड़ी:the cart	0.42	रेलगाड़ी:the train	0.23	है:is	0.36	0.10	525	1297
station	0.52	subway	0.38	train	0.42	0.28	471	

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
स्टेशन:station	0.48	रेलवे:railway	0.45	प्लेटफार्म:platform	0.21	0.28	262	1296
forest	0.44	bamboo	0.29	trees	0.11	0.54	886	
है:is	0.25	के:of	0.28	जंगल:forest	0.11	0.00	981	1288
cars	0.69	parked	0.22	car	0.13	0.59	862	
गाड़ियां:carts	0.50	खड़ी:steep	0.43	है:is	0.40	0.45	1119	1223
church	0.64	cathedral	0.12	a	0.42	0.42	746	
गिरजाघर:the cathedral	0.27	चर्च:the church	0.20	है:is	0.25	0.74	687	1213
kitchen	0.68	stove	0.10	a	0.39	0.38	680	
रसोईघर:kitchen	0.22	रसोई:kitchen	0.26	है:is	0.40	0.65	686	1175
bird	0.51	birds	0.18	a	0.42	0.42	261	
पक्षी:the bird	0.29	चिड़िया:bird	0.22	है:is	0.24	0.54	378	1171
blue	0.70	a	0.36	is	0.21	0.29	2242	
नीले:blue	0.56	रंग:colour	0.65	है:is	0.27	0.26	809	1167
shower	0.55	bathroom	0.24	a	0.46	0.55	583	
नहाने:bathing	0.30	एक:one	0.27	है:is	0.18	0.66	454	1137
fish	0.33	underwater	0.17	coral	0.14	0.38	592	
अंदर:inside	0.26	पानी:water	0.23	समुद्री:marine	0.14	0.02	494	1099
laundromat	0.25	laundry	0.17	dryer	0.14	0.91	608	
मशीन:machine	0.51	है:is	0.26	मशीनें:machines	0.11	0.66	1420	1091
man	0.76	a	0.63	standing	0.10	0.35	2803	
आदमी:man	0.41	एक:one	0.58	यहां:here	0.11	0.26	3435	1068
people	0.75	of	0.41	several	0.11	0.13	1238	
मैदान:field	0.35	लोग:the people	0.38	सारे:all	0.27	0.27	3148	1061
mountains	0.36	mountain	0.38	snow	0.12	0.24	888	
पहाड़:the mountain	0.43	है:is	0.32	दिखाई:visible	0.17	0.21	1469	1053
chairs	0.77	tables	0.13	and	0.26	0.25	860	
कुर्सियां:chairs	0.43	बैठने:sit down	0.17	खुशियां:happiness	0.12	0.43	1522	1044
wooden	0.49	wood	0.17	a	0.40	0.42	1858	
लकड़ी:the wood	0.86	एक:one	0.24	का:of	0.26	0.36	1007	1017
house	0.78	a	0.40	with	0.13	0.14	607	
मकान:house	0.30	है:is	0.52	सामने:front	0.19	0.07	729	1008
truck	0.44	fire	0.46	a	0.43	0.33	481	
गाड़ी:the cart	0.68	है:is	0.41	गाड़ियां:carts	0.13	0.08	265	998
field	0.60	a	0.60	large	0.15	0.35	1242	
घास:grass	0.27	है:is	0.37	हरी:green	0.21	0.34	2340	995

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
table	0.61	tables	0.15	a	0.38	0.42	2418	
टेबल:table	0.24	है:is	0.26	मेज:the table	0.16	0.22	2191	994
children	0.57	kids	0.08	children's	0.05	0.52	723	
बच्चे:children	0.60	बच्चों:the children	0.16	छोटे:small	0.19	0.50	1406	982
statue	0.47	statues	0.21	sculpture	0.08	0.69	1071	
मूर्ति:cculpture	0.50	है:is	0.32	मूर्तियां:sculptures	0.17	0.40	472	974
desert	0.52	sand	0.21	in	0.25	0.59	798	
रेगिस्तान:desert	0.20	है:is	0.33	एक:one	0.21	0.56	591	955
dining	0.49	tables	0.23	room	0.27	0.63	1005	
टेबल:table	0.53	कुर्सियां:chairs	0.18	डाइनिंग:dining	0.16	0.10	450	949
people	0.89	standing	0.09	there	0.20	0.31	2683	
कुछ:some	0.57	लोग:the people	0.48	यहां:here	0.12	0.03	916	943
chairs	0.44	chair	0.20	tables	0.15	0.32	1696	
कुर्सियां:chairs	0.29	कुर्सी:chair	0.25	है:is	0.29	0.50	2851	888
restaurant	0.61	tables	0.13	a	0.43	0.40	662	
रेस्टोरेंट:restaurant	0.32	एक:one	0.33	है:is	0.25	0.77	1443	886
dancing	0.52	dance	0.19	competition	0.08	0.61	310	
हैं:are there	0.34	पहने:to wear	0.22	है:is	0.25	0.01	466	869
forest	0.51	trees	0.10	a	0.43	0.52	821	
जंगल:forest	0.48	है:is	0.30	एक:one	0.22	0.52	1198	868
sitting	0.63	people	0.50	table	0.08	0.26	1509	
बैठे:sitting	0.64	हैं:are there	0.53	लोग:the people	0.36	0.33	1183	862
child	0.31	boy	0.27	young	0.23	0.41	737	
बच्चा:child	0.34	एक:one	0.52	बच्चे:children	0.22	0.41	611	858
microphone	0.43	stage	0.14	microphones	0.07	0.62	445	
मंच:forum	0.30	माइक:mike	0.27	है:is	0.34	0.30	327	841
beach	0.43	a	0.33	cliff	0.10	0.34	634	
किनारे:edge	0.50	समुद्र:sea	0.42	के:of	0.56	0.12	542	823
rocks	0.22	rock	0.18	rocky	0.12	0.48	1916	
पत्थर:stone	0.59	है:is	0.30	हैं:are there	0.15	0.28	847	819
dirt	0.33	construction	0.18	and	0.16	0.41	1149	
काम:work	0.58	निर्माण:construction	0.25	रहा:stayed	0.35	0.18	415	805
ocean	0.55	the	0.55	beach	0.16	0.48	1012	
किनारे:edge	0.42	समंदर:ocean	0.27	समुंदर:saltwater	0.27	0.20	1074	803
clouds	0.52	sky	0.20	cloudy	0.18	0.47	1002	

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
बादल:cloud	0.75	आसमान:sky	0.42	है:is	0.41	0.42	928	801
shop	0.37	store	0.13	storefront	0.05	0.47	767	
दुकान:shop	0.33	है:is	0.29	दृश्य:view	0.12	0.18	1229	800
trees	0.55	tree	0.27	and	0.16	0.24	3187	
पेड़:the trees	0.69	है:is	0.41	हैं:are there	0.20	0.13	1610	795
airplane	0.19	plane	0.14	playground	0.08	0.64	1501	
हवाई:airy	0.52	जहाज:ship	0.46	है:is	0.17	0.59	686	782
water	0.78	body	0.45	of	0.56	0.22	1838	
पानी:water	0.51	है:is	0.33	नदी:river	0.23	0.27	2828	779
street	0.65	a	0.36	the	0.43	0.42	978	
सड़क:road	0.38	है:is	0.30	पीले:yellow	0.24	0.23	1850	778
trees	0.79	and	0.21	with	0.13	0.19	1634	
पेड़:the trees	0.49	पेड़ों:the trees	0.26	कुछ:some	0.37	0.09	1396	770
sunset	0.32	sun	0.26	setting	0.11	0.74	710	
आसमान:sky	0.69	है:is	0.34	में:in	0.25	0.21	1523	749
door	0.47	doors	0.23	the	0.32	0.35	1000	
दरवाजा:door	0.51	दरवाजे:the doors	0.26	है:is	0.37	0.27	435	746
grass	0.57	green	0.25	grassy	0.12	0.32	1273	
घास:grass	0.36	है:is	0.32	मैदान:field	0.20	0.40	2306	728
woman	0.57	a	0.54	is	0.20	0.37	2117	
एक:one	0.75	औरत:the woman	0.42	लड़की:girl	0.27	0.03	1989	725
building	0.79	buildings	0.07	the	0.23	0.08	846	
इमारत:the building	0.40	बड़ा:big	0.37	एक:one	0.40	0.15	1036	710
water	0.60	ocean	0.20	the	0.42	0.16	1675	
समंदर:ocean	0.33	समुद्र:sea	0.28	समुंदर:saltwater	0.20	0.32	930	699
woman	0.64	a	0.52	women	0.10	0.27	1455	
औरत:the woman	0.47	एक:one	0.42	है:is	0.26	0.10	618	679
rocks	0.25	rock	0.32	formation	0.13	0.33	1283	
पहाड़:the mountain	0.37	पहाड़ी:hill	0.27	एक:one	0.21	0.24	1832	669
people	0.83	walking	0.20	there	0.23	0.09	773	
लोग:the people	0.69	कुछ:some	0.73	दिखाई:visible	0.16	0.19	2058	639
two	0.51	people	0.28	men	0.28	0.11	1045	
दो:two	0.52	यहां:here	0.14	आदमी:man	0.18	0.11	950	633
house	0.66	a	0.41	houses	0.07	0.15	710	
घर:home	0.48	मकान:house	0.21	है:is	0.30	0.14	1389	628

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
wall	0.65	on	0.40	the	0.50	0.33	1431	
दीवार:wall	0.64	पर:on	0.47	है:is	0.28	0.23	873	616
red	0.77	and	0.22	a	0.29	0.15	1042	
लाल:red	0.78	रंग:colour	0.68	है:is	0.36	0.27	1135	597
walls	0.34	wall	0.33	the	0.59	0.24	667	
कमरे:the room	0.57	है:is	0.36	कमरा:rooms	0.18	0.21	1625	597
clothing	0.24	clothes	0.14	shirt	0.11	0.44	658	
कपड़े:dresses	0.51	सारे:all	0.14	है:is	0.16	0.18	649	593
yellow	0.72	and	0.20	is	0.18	0.36	1220	
पीले:yellow	0.51	रंग:colour	0.52	है:is	0.33	0.31	870	586
road	0.25	walkway	0.11	path	0.15	0.14	799	
रास्ता:way	0.42	दोनों:both	0.19	है:is	0.27	0.23	674	544
building	0.80	large	0.11	a	0.34	0.17	1855	
बिल्डिंग:building	0.59	दिखाई:visible	0.16	विशाल:vishal	0.10	0.21	664	526
shelves	0.31	shelf	0.22	on	0.30	0.49	551	
सारे:all	0.41	बहुत:very	0.46	यहां:here	0.11	0.08	1439	471
cream	0.40	ice	0.47	of	0.20	0.50	956	
खाने:food	0.46	सामान:luggage	0.16	रखी:kept	0.22	0.35	1325	454
girl	0.54	girls	0.21	young	0.17	0.50	737	
बच्चे:children	0.31	है:is	0.23	लड़की:girl	0.11	0.24	1147	449
river	0.66	a	0.47	creek	0.09	0.54	1007	
नदी:river	0.48	है:is	0.35	एक:one	0.44	0.16	777	428
building	0.53	buildings	0.17	street	0.09	0.29	4499	
मकान:house	0.30	है:is	0.41	कुर्सियां:chairs	0.18	0.39	4098	381
water	0.63	waterfall	0.06	the	0.35	0.19	1983	
सफेद:white	0.33	पानी:water	0.35	है:is	0.35	0.07	1417	366
plants	0.41	plant	0.16	green	0.16	0.45	1309	
फूल:flower	0.56	है:is	0.26	हैं:are there	0.20	0.26	849	351
station	0.37	telephone	0.25	gas	0.38	0.28	706	
इमारतों:buildings	0.29	हैं:are there	0.42	मारते:kills	0.20	0.51	804	95
tracks	0.38	train	0.42	track	0.23	0.67	947	
सीढ़ियां:stairs	0.25	चिड़िया:bird	0.14	है:is	0.28	0.81	1305	32
iceberg	0.22	ice	0.29	glacier	0.16	0.77	378	
आदमी:man	0.66	एक:one	0.78	यहां:here	0.14	0.26	2143	16
baseball	0.83	a	0.55	stadium	0.08	0.64	659	

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
इमारत:the building	0.27	तस्वीर:picture	0.21	बड़ा:big	0.25	0.15	1511	12
desk	0.32	books	0.14	a	0.37	0.65	1941	
मैदान:field	0.36	घास:grass	0.42	यहां:here	0.13	0.12	1149	3
store	0.68	a	0.30	clothing	0.07	0.39	797	
						0.00	0	0
cemetery	0.58	graveyard	0.10	a	0.54	0.80	255	
						0.00	0	0
tower	0.56	water	0.22	towers	0.07	0.69	1055	
						0.00	0	0
this	0.12	a	0.32	there	0.12	0.11	8245	
						0.00	0	0
trees	0.77	green	0.14	and	0.17	0.38	3348	
						0.00	0	0
stone	0.69	a	0.35	of	0.23	0.27	1131	
						0.00	0	0
						0.00	0	
चट्टान:rock	0.22	चट्टानों:rocks	0.17	है:is	0.26	0.49	813	0
photograph	0.08	a	0.41	this	0.14	0.11	7449	
						0.00	0	0
black	0.76	a	0.37	wearing	0.13	0.17	1264	
						0.00	0	0
						0.00	0	
कमरे:the room	0.79	में:in	0.51	एक:one	0.50	0.31	1646	0
						0.00	0	
लोग:the people	0.62	सारे:all	0.44	बहुत:very	0.29	0.17	1938	0
brick	0.77	building	0.36	red	0.17	0.51	1410	
						0.00	0	0
a	0.26	the	0.21	of	0.15	0.24	66584	
						0.00	0	0
boat	0.60	a	0.53	boats	0.10	0.21	384	
						0.00	0	0
sky	0.79	blue	0.52	the	0.42	0.12	474	
						0.00	0	0
classroom	0.47	room	0.34	locker	0.12	0.66	585	
						0.00	0	0

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
there	0.08	photograph	0.04	a	0.16	0.02	1539	
						0.00	0	0
green	0.57	and	0.16	ground	0.06	0.11	1061	
						0.00	0	0
a	0.28	of	0.16	the	0.18	0.26	55201	
						0.00	0	0
standing	0.51	people	0.12	wearing	0.11	0.21	1375	
						0.00	0	0
church	0.66	a	0.29	chapel	0.06	0.41	718	
						0.00	0	0
large	0.56	structure	0.24	a	0.48	0.06	807	
						0.00	0	0
clouds	0.56	sky	0.17	desert	0.16	0.11	194	
						0.00	0	0
photo	0.76	close	0.80	showcasing	0.19	0.04	230	
						0.00	0	0
hospital	0.40	crib	0.21	bed	0.17	0.62	556	
						0.00	0	0
red	0.73	a	0.35	is	0.20	0.14	947	
						0.00	0	0
archway	0.46	$spoken_noise$	0.17	stone	0.14	0.37	287	
						0.00	0	0
walkway	0.18	hallway	0.16	white	0.18	0.21	705	
						0.00	0	0
house	0.71	houses	0.06	a	0.32	0.25	1226	
						0.00	0	0
walking	0.77	people	0.30	down	0.16	0.40	934	
						0.00	0	0
night	0.37	lights	0.24	time	0.21	0.36	635	
						0.00	0	0
						0.00	0	
कमरे:the room	0.82	अंदर:inside	0.47	एक:one	0.37	0.30	1658	0
background	0.82	in	0.85	the	0.85	0.30	1860	
						0.00	0	0
fountain	0.63	water	0.21	a	0.46	0.65	550	

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
						0.00	0	0
sky	0.67	blue	0.41	the	0.46	0.24	973	
						0.00	0	0
brick	0.59	bricks	0.12	a	0.30	0.14	425	
						0.00	0	0
structure	0.56	construction	0.16	structures	0.14	0.38	991	
						0.00	0	0
inside	0.65	photograph	0.18	taken	0.14	0.33	1830	
						0.00	0	0
door	0.36	doors	0.20	the	0.35	0.16	453	
						0.00	0	0
locker	0.40	lockers	0.21	room	0.29	0.73	403	
						0.00	0	0
white	0.34	a	0.33	and	0.14	0.15	3398	
						0.00	0	0
						0.00	0	
है:is	0.26	एक:one	0.19	यहां:here	0.06	0.16	32659	0
						0.00	0	
पहाड़:the mountain	0.50	है:is	0.40	पहाड़ी:hill	0.14	0.09	522	0
						0.00	0	
आसमान:sky	0.74	नीला:blue	0.57	ऊपर:up	0.41	0.21	1359	0
						0.00	0	
पहाड़:the mountain	0.50	है:is	0.40	पहाड़ी:hill	0.14	0.09	522	0
						0.00	0	
कमरे:the room	0.79	में:in	0.51	एक:one	0.50	0.31	1646	0
						0.00	0	
माइक:mike	0.20	एक:one	0.25	है:is	0.25	0.24	241	0
						0.00	0	
लकड़ी:the wood	0.75	एक:one	0.36	यहां:here	0.11	0.17	483	0
						0.00	0	
हरे:green	0.34	भरे:fill up	0.22	एक:one	0.15	0.14	1206	0
						0.00	0	
आसमान:sky	0.64	है:is	0.34	और:and	0.26	0.31	2354	0
						0.00	0	
है:is	0.25	एक:one	0.16	हैं:are there	0.08	0.04	7172	0

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
						0.00	0	
पौधे:plants	0.55	पेड़:the trees	0.53	दिखाई:visible	0.32	0.46	3023	0
						0.00	0	
दुकान:shop	0.88	है:is	0.50	एक:one	0.30	0.26	684	0
						0.00	0	
छोटा:small	0.24	एक:one	0.37	झोपड़ी:hut	0.11	0.12	1078	0
						0.00	0	
आसमान:sky	0.76	ऊपर:up	0.35	है:is	0.44	0.16	1031	0
						0.00	0	
पत्थर:stone	0.62	है:is	0.24	पत्थरों:the stones	0.12	0.33	967	0
						0.00	0	
आसमान:sky	0.65	है:is	0.79	नीला:blue	0.67	0.05	403	0
						0.00	0	
लिखा:written	0.63	हुआ:happened	0.32	है:is	0.30	0.41	1222	0
						0.00	0	
है:is	0.52	और:and	0.24	रही:doing	0.35	0.01	765	0
						0.00	0	
यह:this	0.03	एक:one	0.06	यहां:here	0.00	0.00	527	0
						0.00	0	
लिखा:written	0.81	है:is	0.80	हुआ:happened	0.54	0.35	866	0
						0.00	0	
है:is	0.32	और:and	0.24	नजर:vision	0.24	0.00	266	0
						0.00	0	
है:is	0.45	हैं:are there	0.14	और:and	0.08	0.02	3091	0
						0.00	0	
सूखे:dry	0.38	पेड़:the trees	0.39	नजर:vision	0.25	0.48	677	0
						0.00	0	
मैदान:field	0.32	खेत:farm	0.36	है:is	0.27	0.06	736	0
						0.00	0	
पेड़:the trees	0.64	है:is	0.36	नजर:vision	0.35	0.08	1014	0
						0.00	0	
इमारतों:buildings	0.29	हैं:are there	0.42	मारते:kills	0.20	0.51	804	0
walking	0.76	people	0.39	man	0.18	0.21	465	
						0.00	0	0
						0.00	0	

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
है:is	0.39	दिखाई:visible	0.20	इमारत:the building	0.09	0.00	663	0
खूबसूरत:beautiful	0.77	एक:one	0.30	यहां:here	0.09	0.00 0.21	0 388	0
सफेद:white	0.85	रंग:colour	0.71	है:is	0.30	0.00 0.15	0 1319	0
इमारत:the building	0.34	जिसके:whose	0.42	है:is	0.56	0.00 0.08	0 735	0
है:is	0.28	एक:one	0.13	और:and	0.11	0.00 0.08	0 19848	0
है:is	0.24	एक:one	0.23	यहां:here	0.07	0.00 0.10	0 17641	0
कुर्सियां:chairs	0.43	है:is	0.47	कुर्सी:chair	0.21	0.00 0.44	0 1709	0
दीवार:wall	0.26	है:is	0.24	लाल:red	0.20	0.00 0.07	0 584	0
लोहे:iron	0.41	है:is	0.36	सीढ़ियां:stairs	0.17	0.00 0.16	0 327	0
सफेद:white	0.85	रंग:colour	0.71	है:is	0.30	0.00 0.15	0 1319	0
है:is	0.28	एक:one	0.15	और:and	0.10	0.00 0.15	0 34384	0
मशीन:machine	0.46	है:is	0.29	एक:one	0.21	0.00 0.37	0 809	0
कंप्यूटर:computer	0.45	लैपटॉप:laptop	0.30	है:is	0.27	0.00 0.48	0 458	0
इमारत:the building	0.23	है:is	0.36	दिखाई:visible	0.23	0.00 0.17	0 2176	0
है:is	0.45	झंडा:flag	0.26	हैं:are there	0.38	0.00 0.00	0 391	0
समुद्र:sea	0.48	है:is	0.37	समंदर:ocean	0.18	0.00 0.17	0 527	0
कमरे:the room	0.82	अंदर:inside	0.47	एक:one	0.37	0.00 0.30	0 1658	0
है:is	0.26	एक:one	0.19	यहां:here	0.06	0.00 0.16	0 32659	0

Continued on next page

Table A.1 – *Continued from previous page*

Word 1	p1	Word 2	p2	Word 3	p3	Coverage	N	Similarity
						0.00	0	
कुर्सियां:chairs	0.32	है:is	0.44	कुर्सी:chair	0.24	0.13	652	0
						0.00	0	
पेड़:the trees	0.66	है:is	0.32	और:and	0.28	0.12	1482	0
						0.00	0	
रोशनी:light	0.28	रात:night	0.36	है:is	0.40	0.17	418	0
shoes	0.55	store	0.11	shoe	0.16	0.57	759	
पेड़:the trees	0.69	है:is	0.41	हैं:are there	0.20	0.13	1610	-4
bridge	0.84	over	0.12	a	0.33	0.31	700	
आसमान:sky	0.74	नीला:blue	0.57	ऊपर:up	0.41	0.21	1359	-37

Table A.2: Trilingual word clusters with each block representing English, Hindi and Japanese clusters within a meta-cluster. $p1$, $p2$ and $p3$ refer to the purity scores of the top three words in the cluster. N is the number of peaks in the meta-cluster for each language.

Word 1	p1	Word 2	p2	Word 3	p3	Coverage	N	Similarity
kitchen	0.89	a	0.40	the	0.25	0.79	1098	
रसोईघर:kitchen	0.23	रसोई:kitchen	0.25	है:is	0.34	0.88	886	2158
キッチン:kitchen	0.73	の:of	0.35	ダイニング:dining	0.03	0.70	1396	
bus	0.70	inside	0.16	a	0.48	0.57	353	
बस:bus	0.80	अंदर:inside	0.32	एक:one	0.42	0.31	250	1945
バス:bus	0.77	の:of	0.70	車内:Inside the car	0.21	0.53	374	
windmill	0.38	windmills	0.21	wind	0.21	0.83	606	
पवन:air	0.40	दिखाई:visible	0.22	चक्की:mill	0.20	0.75	464	1697
風車:Windmill	0.48	風力:Wind power	0.22	の:of	0.34	0.70	819	
boxing	0.59	ring	0.29	boxers	0.13	0.75	434	
बॉक्सिंग:boxing	0.71	रिंग:ring	0.26	एक:one	0.23	0.66	263	1566
ボクシング:boxing	0.54	リング:ring	0.30	の:of	0.48	0.89	491	
black	0.85	white	0.82	photo	0.32	0.22	1383	
तस्वीर:picture	0.48	ब्लैक:the black	0.32	यह:this	0.26	0.05	496	1542
白黒:Black and white	0.62	モノクロ:Monochrome	0.22	写真:Photo	0.50	0.91	2065	
horse	0.50	horses	0.30	a	0.39	0.70	627	
घोड़े:the horse	0.32	घोड़ा:horse	0.12	है:is	0.21	0.76	696	1524
馬:Horse	0.55	競馬:Horse racing	0.13	の:of	0.33	0.49	707	
bicycle	0.20	motorcycle	0.17	bike	0.14	0.77	459	
साइकिल:bicycle	0.34	मोटरसाइकिल:motorcycle	0.09	है:is	0.23	0.78	603	1512

Continued on next page

Table A.2 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
マイク:Microphone	0.20	自転:rotation	0.20	バイク:bike	0.17	0.66	1129	
bedroom	0.30	bed	0.41	a	0.42	0.85	1241	
बिस्तर:bed	0.62	है:is	0.34	सोने:the gold	0.21	0.73	847	1287
ベッド:bed	0.63	の:of	0.35	が:But	0.30	0.51	1917	
statue	0.44	statues	0.17	sculpture	0.11	0.75	902	
मूर्ति:cculpture	0.47	मूर्तियां:sculptures	0.24	है:is	0.34	0.69	529	1247
銅像:Bronze statue	0.12	の:of	0.45	が:But	0.33	0.76	1789	
fish	0.54	station	0.26	gas	0.28	0.65	731	
मछली:fish	0.47	मछलियां:fish	0.23	है:is	0.21	0.63	469	1212
ガソリン:gasoline	0.23	魚:fish	0.29	水槽:Water tank	0.13	0.63	1317	
staircase	0.25	stairs	0.26	steps	0.15	0.84	1135	
सीढ़ियां:stairs	0.27	चिड़िया:bird	0.14	है:is	0.27	0.93	1342	1134
階段:Stairs	0.61	の:of	0.33	が:But	0.24	0.64	2285	
asian	0.43	chinese	0.21	building	0.17	0.39	588	
मंदिर:temple	0.47	है:is	0.24	एक:one	0.21	0.38	453	1107
五重塔:5-story pagoda	0.09	の:of	0.56	寺:temple	0.18	0.83	710	
boat	0.51	boats	0.20	a	0.39	0.66	1396	
नाव:the boat	0.28	है:is	0.29	जहाज:ship	0.16	0.76	1225	1092
船:boat	0.36	ボート:boat	0.20	の:of	0.35	0.58	2175	
baseball	0.42	course	0.21	golf	0.27	0.76	1481	
खिलाड़ी:the player	0.42	मैदान:field	0.19	यहां:here	0.11	0.75	1107	997
ゴルフ:golf	0.30	野球:baseball	0.28	の:of	0.44	0.68	1966	
children	0.26	child	0.12	young	0.17	0.76	2088	
बच्चे:children	0.43	बच्चा:child	0.16	छोटे:small	0.17	0.58	2003	926
子供:children	0.32	男の子:boy	0.13	女の子:girl	0.14	0.62	3287	
ocean	0.53	beach	0.35	the	0.57	0.79	1661	
समुद्र:sea	0.28	समंदर:ocean	0.25	किनारे:edge	0.26	0.66	2882	876
海:Sea	0.48	砂浜:Sandy beach	0.08	海岸:Coast	0.06	0.64	4676	
couch	0.67	couches	0.23	white	0.12	0.51	319	
सोफा:sofa	0.27	है:is	0.29	सोफे:sofa	0.20	0.80	939	832
ソファー:sofa	0.59	が:But	0.32	の:of	0.34	0.60	1558	
bowling	0.89	alley	0.63	a	0.54	0.55	236	
						0.00	0	812
ボーリング:Bowling	0.80	の:of	0.36	場:Place	0.46	0.66	459	
shower	0.86	a	0.39	head	0.12	0.55	295	
						0.00	0	810
シャワー:shower	0.72	ルーム:Room	0.32	の:of	0.27	0.56	423	

Continued on next page

Table A.2 – *Continued from previous page*

Word 1	p1	Word 2	p2	Word 3	p3	Coverage	N	Similarity
bridge	0.58	archway	0.10	stone	0.09	0.71	2216	
है:is	0.37	पुल:the bridge	0.33	एक:one	0.36	0.01	1368	796
橋:bridge	0.37	吊り橋:suspension bridge	0.12	の:of	0.32	0.56	2502	
books	0.44	bookstore	0.15	bookshelf	0.10	0.76	726	
सामान:luggage	0.41	किताबें:the books	0.29	बहुत:very	0.25	0.30	1029	771
本棚:Bookshelf	0.30	本:Book	0.27	本屋:bookstore	0.09	0.67	933	
fire	0.69	truck	0.26	fireplace	0.09	0.28	378	
						0.00	0	754
消防:Firefighting	0.76	車:car	0.45	の:of	0.40	0.52	362	
airplane	0.33	airport	0.20	plane	0.16	0.70	712	
हवाई:airy	0.89	जहाज:ship	0.86	एक:one	0.24	0.64	379	751
天井:ceiling	0.27	飛行:Flight	0.26	空港:airport	0.09	0.36	1966	
telephone	0.56	booth	0.52	phone	0.27	0.58	306	
						0.00	0	745
電話:phone	0.73	ボックス:box	0.51	公衆:public	0.21	0.39	448	
construction	0.35	machine	0.14	laundromat	0.08	0.73	1269	
मशीन:machine	0.56	मशीनें:machines	0.12	दिखाई:visible	0.17	0.73	1221	722
工事:Construction	0.13	洗濯:Washing	0.10	の:of	0.37	0.74	2898	
train	0.60	station	0.15	tracks	0.15	0.58	1999	
रेल:rail	0.32	पटरी:track	0.23	रेलगाड़ी:the train	0.14	0.66	1083	705
線路:line	0.18	電車:Electric train	0.20	駅:station	0.13	0.65	3226	
flag	0.49	american	0.26	flags	0.17	0.70	717	
						0.00	0	687
アメリカ:America	0.50	国旗:Flag	0.49	の:of	0.56	0.41	345	
flowers	0.70	flower	0.15	with	0.15	0.56	1045	
फूल:flower	0.63	रंग:colour	0.30	है:is	0.21	0.56	1491	678
花:flower	0.55	の:of	0.39	白い:white	0.07	0.50	2717	
chairs	0.69	chair	0.16	table	0.10	0.45	1634	
कुर्सियां:chairs	0.43	कुर्सी:chair	0.20	है:is	0.24	0.67	2429	653
椅子:Chair	0.53	が:But	0.38	の:of	0.37	0.33	2213	
yellow	0.80	and	0.24	a	0.27	0.49	1459	
पीले:yellow	0.61	रंग:colour	0.62	है:is	0.29	0.37	858	617
黄色い:yellow	0.34	黄色:yellow	0.21	の:of	0.24	0.67	3209	
red	0.77	a	0.33	and	0.23	0.55	3461	
लाल:red	0.72	रंग:colour	0.57	है:is	0.26	0.65	2793	615
赤い:red	0.32	赤色:red	0.08	オレンジ:Orange	0.07	0.74	8258	
people	0.89	standing	0.09	several	0.10	0.51	4088	

Continued on next page

Table A.2 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
लोग:the people	0.66	कुछ:some	0.53	सारे:all	0.20	0.46	4933	612
たくさん:A lot	0.22	人:Man	0.36	の:of	0.41	0.22	7212	
black	0.82	a	0.39	wearing	0.15	0.13	784	
काले:the black	0.78	रंग:colour	0.73	है:is	0.27	0.29	987	608
黒い:black	0.47	黒色:Black	0.07	黒:black	0.07	0.38	2852	
house	0.66	houses	0.11	a	0.36	0.38	1806	
मकान:house	0.40	है:is	0.36	घर:home	0.32	0.53	3947	607
家:House	0.31	屋根:roof	0.17	の:of	0.42	0.31	4091	
people	0.46	two	0.47	men	0.32	0.05	761	
दो:two	0.53	यहां:here	0.15	लोग:the people	0.32	0.12	847	600
二人:Two persons	0.36	男性:male	0.19	人:Man	0.23	0.30	1691	
mountain	0.39	mountains	0.32	background	0.07	0.56	2774	
पहाड़:the mountain	0.46	है:is	0.34	पहाड़ी:hill	0.17	0.68	4548	597
山:Mountain	0.50	岩山:Iwayama	0.09	の:of	0.34	0.61	5453	
store	0.30	shop	0.12	a	0.35	0.59	2305	
दुकान:shop	0.78	है:is	0.35	एक:one	0.36	0.58	1531	597
店:shop	0.26	の:of	0.50	お:The	0.16	0.34	2916	
woman	0.64	a	0.54	women	0.10	0.65	3250	
औरत:the woman	0.44	एक:one	0.61	लड़की:girl	0.17	0.59	3759	592
女性:Woman	0.62	が:But	0.32	の:of	0.35	0.51	5711	
cars	0.30	car	0.27	parking	0.18	0.65	1909	
गाड़ी:the cart	0.41	गाड़ियां:carts	0.30	है:is	0.32	0.56	3055	587
車:car	0.59	が:But	0.41	の:of	0.34	0.40	2859	
night	0.22	sunset	0.11	lights	0.16	0.72	1871	
है:is	0.49	रोशनी:light	0.23	रही:doing	0.33	0.02	2132	587
夜:Night	0.11	ライト:Light	0.11	照明:illumination	0.08	0.77	4402	
rocks	0.37	rock	0.36	formation	0.10	0.63	1651	
पत्थर:stone	0.40	है:is	0.27	यहां:here	0.08	0.63	2629	587
岩:rock	0.28	石:stone	0.12	大きな:big	0.09	0.67	2985	
room	0.60	classroom	0.23	living	0.27	0.24	1368	
कमरे:the room	0.66	कमरा:rooms	0.20	एक:one	0.40	0.39	2470	587
部屋:room	0.48	室内:Indoor	0.19	の:of	0.59	0.40	3095	
street	0.25	road	0.25	walking	0.15	0.49	2799	
सड़क:road	0.52	है:is	0.34	रास्ता:way	0.21	0.58	3370	578
道路:road	0.36	道:road	0.22	の:of	0.35	0.49	3821	
table	0.44	tables	0.22	restaurant	0.16	0.50	3556	
टेबल:table	0.39	है:is	0.27	और:and	0.25	0.42	2220	564

Continued on next page

Table A.2 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
テーブル:table	0.44	レストラン:Restaurant	0.09	机:desk	0.11	0.45	5292	
wooden	0.43	wood	0.26	of	0.24	0.24	1112	
लकड़ी:the wood	0.89	का:of	0.34	बना:make	0.18	0.44	1118	557
木:wood	0.43	木製:wooden	0.11	の:of	0.37	0.07	1178	
green	0.73	and	0.17	a	0.27	0.26	1922	
हरे:green	0.42	मैदान:field	0.15	रंग:colour	0.33	0.24	1737	540
緑色:green	0.39	緑:Green	0.21	の:of	0.34	0.57	3542	
windows	0.37	window	0.27	glass	0.18	0.62	2311	
शीशे:glass	0.28	खिड़की:window	0.25	है:is	0.38	0.56	940	535
窓:window	0.27	ガラス:Glass	0.19	の:of	0.33	0.53	4527	
water	0.44	body	0.20	pool	0.11	0.47	6216	
पानी:water	0.52	है:is	0.28	नदी:river	0.19	0.59	5815	532
川:river	0.18	湖:lake	0.07	水:water	0.08	0.53	9600	
buildings	0.35	city	0.33	skyscrapers	0.07	0.43	1640	
मारते:kills	0.23	इमारतों:buildings	0.18	हैं:are there	0.22	0.57	1348	524
ビル:building	0.38	高層:High-rise	0.18	街並み:Cityscape	0.07	0.54	1891	
grass	0.35	field	0.35	green	0.19	0.48	2762	
घास:grass	0.35	है:is	0.28	यहां:here	0.09	0.76	3817	523
芝生:lawn	0.27	草:grass	0.13	畑:field	0.09	0.71	6510	
bird	0.59	birds	0.29	a	0.41	0.51	230	
						0.00	0	501
鳥:bird	0.58	の:of	0.38	一:one	0.10	0.42	480	
shoes	0.67	of	0.34	shoe	0.12	0.55	436	
						0.00	0	492
靴:shoes	0.40	靴屋:Shoe store	0.13	の:of	0.35	0.48	714	
stone	0.37	church	0.33	castle	0.13	0.33	2238	
गिरजाघर:the cathedral	0.16	है:is	0.28	पुरानी:old	0.15	0.86	1345	480
石造り:Stone	0.12	境界:boundary	0.12	石:stone	0.11	0.64	3447	
blue	0.78	sky	0.10	a	0.36	0.30	2004	
नीले:blue	0.51	रंग:colour	0.55	है:is	0.32	0.49	1610	479
青い:blue	0.33	青色:Blue	0.11	水色:light blue	0.10	0.51	3261	
desk	0.40	computer	0.21	office	0.18	0.64	1451	
						0.00	0	475
パソコン:computer	0.60	ノート:Note	0.16	が:But	0.27	0.73	761	
snow	0.36	snowy	0.10	stage	0.08	0.79	4744	
बर्फ:ice	0.40	है:is	0.21	खाने:food	0.12	0.72	3882	467
雪:snow	0.28	雪山:Snowy mountain	0.11	スキー:Ski	0.08	0.65	6141	

Continued on next page

Table A.2 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
building	0.77	large	0.12	buildings	0.08	0.14	1448	
इमारत:the building	0.30	बिल्डिंग:building	0.22	दिखाई:visible	0.22	0.33	2998	440
建物:building	0.69	大きな:big	0.09	の:of	0.44	0.14	2207	
man	0.81	a	0.67	standing	0.10	0.38	2535	
आदमी:man	0.54	एक:one	0.69	व्यक्ति:person	0.18	0.41	3564	437
男性:male	0.54	一人:One person	0.09	が:But	0.27	0.30	3962	
bar	0.67	a	0.46	barn	0.10	0.42	373	
						0.00	0	432
カウンター:counter	0.75	の:of	0.43	が:But	0.24	0.49	695	
wall	0.62	on	0.50	walls	0.16	0.27	1024	
दीवार:wall	0.61	है:is	0.33	पर:on	0.31	0.56	2060	428
壁:wall	0.42	の:of	0.33	白い:white	0.06	0.28	4432	
trees	0.64	tree	0.15	forest	0.06	0.65	6397	
पेड़:the trees	0.59	है:is	0.28	और:and	0.20	0.69	9072	419
木:wood	0.34	森:Woods	0.12	の:of	0.39	0.50	10152	
tower	0.47	lighthouse	0.25	water	0.21	0.56	927	
						0.00	0	415
灯台:Lighthouse	0.15	頭:Head	0.16	の:of	0.41	0.73	1301	
plants	0.24	garden	0.25	bushes	0.17	0.54	2196	
पौधे:plants	0.37	बगीचा:garden	0.13	है:is	0.22	0.33	2912	409
植物:plant	0.38	公園:park	0.12	庭:garden	0.10	0.56	3772	
bamboo	0.28	dancing	0.21	damn	0.09	0.66	710	
						0.00	0	390
竹林:Bamboo forest	0.19	ダンス:dance	0.17	竹:bamboo	0.10	0.85	988	
clouds	0.14	door	0.18	doors	0.09	0.53	3826	
है:is	0.34	बादल:cloud	0.23	दरवाजा:door	0.14	0.03	4455	377
雲:cloud	0.16	空:Sky	0.08	扉:door	0.09	0.52	6340	
brick	0.79	building	0.32	a	0.35	0.60	1477	
						0.00	0	361
レンガ:Brick	0.28	茶色い:Brown	0.19	茶色:Brown	0.13	0.48	1938	
sign	0.66	says	0.19	a	0.42	0.28	532	
						0.00	0	288
看板:Signboard	0.29	書か:Writing	0.19	と:When	0.20	0.62	2927	
desert	0.34	dirt	0.26	sand	0.16	0.76	1518	
						0.00	0	288
砂漠:Desert	0.23	土:soil	0.20	ゴミ:garbage	0.17	0.71	1550	
fence	0.29	tent	0.14	a	0.41	0.34	840	

Continued on next page

Table A.2 – *Continued from previous page*

Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
						0.00	0	285
ベンチ:bench	0.19	ピンク:pink	0.16	テント:tent	0.15	0.64	2649	
						0.00	0	
बैठे:sitting	0.48	लोग:the people	0.38	हैं:are there	0.25	0.42	1761	274
座つ:Sitting	0.61	椅子:Chair	0.15	て:The	0.48	0.40	1290	
a	0.32	of	0.16	the	0.17	0.20	32844	
						0.00	0	0
						0.00	0	
						0.00	0	
यह:this	0.07	एक:one	0.06	□□:mp	0.00	0.01	301	0
						0.00	0	
						0.00	0	
						0.00	0	0
の:of	0.38	が:But	0.17	大きな:big	0.03	0.09	26228	
						0.00	0	
						0.00	0	0
の:of	0.35	が:But	0.19	に:To	0.16	0.17	61558	
						0.00	0	
						0.00	0	0
白い:white	0.47	白色:White	0.10	壁:wall	0.12	0.33	5772	
						0.00	0	
						0.00	0	0
写真:Photo	0.06	の:of	0.18	建物:building	0.03	0.04	3562	
						0.00	0	
						0.00	0	0
空:Sky	0.50	青空:blue sky	0.18	青い:blue	0.06	0.31	2505	
						0.00	0	
सफेद:white	0.83	रंग:colour	0.62	है:is	0.29	0.29	2560	0
						0.00	0	
						0.00	0	
है:is	0.19	हैं:are there	0.07	और:and	0.08	0.00	374	0
						0.00	0	
						0.00	0	
एक:one	0.26	यहां:here	0.09	है:is	0.19	0.16	14810	0
						0.00	0	
						0.00	0	
आसमान:sky	0.82	नीला:blue	0.43	है:is	0.54	0.44	2544	0

Continued on next page

Table A.2 – *Continued from previous page*

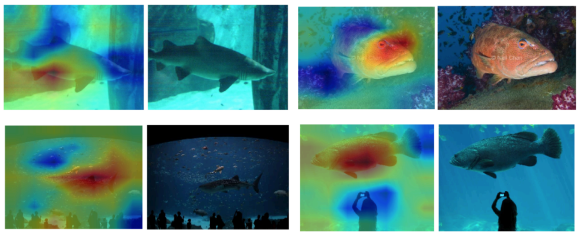
Word 1	**p1**	**Word 2**	**p2**	**Word 3**	**p3**	**Coverage**	**N**	Similarity
						0.00	0	
white	0.77	building	0.09	a	0.33	0.19	2229	
						0.00	0	0
						0.00	0	
						0.00	0	
है:is	0.26	और:and	0.14	एक:one	0.12	0.12	19955	0
						0.00	0	
						0.00	0	
है:is	0.23	एक:one	0.18	यहां:here	0.06	0.19	34771	0
						0.00	0	
sky	0.73	blue	0.46	the	0.46	0.13	472	
						0.00	0	0
						0.00	0	
and	0.03	the	0.06	on	0.03	0.02	2223	
						0.00	0	0
						0.00	0	
a	0.24	the	0.20	and	0.11	0.21	39802	
						0.00	0	0
						0.00	0	
						0.00	0	
						0.00	0	0
の:of	0.33	が:But	0.20	に:To	0.20	0.10	32997	

Appendix B

Figures

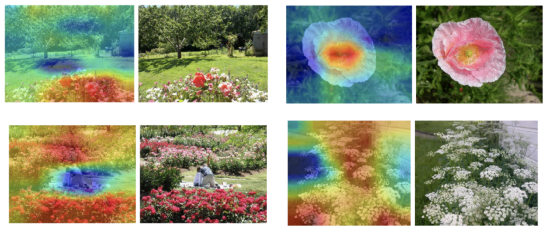
Figure B-1: Picture dictionary representing four-way agreement between English speech caption, Hindi speech caption, Japanese speech caption and Image pixels. We present the text transcriptions of the clustered speech segments with their corresponding cluster purities.

ocean (0.53), समुद्र:sea (0.53), 海:Sea (0.48)

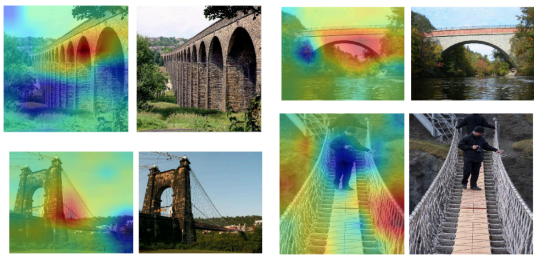


fish (0.75), मछली:fish (0.47), 魚:fish (0.40)

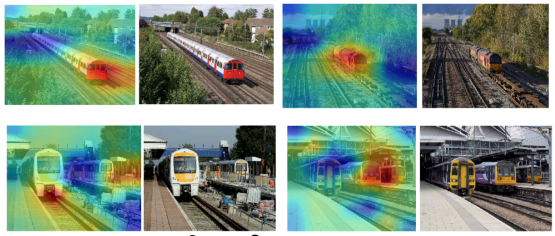
horse(s) (0.8), घोड़े:the horse (0.44), 馬:Horse (0.68)



flower(s) (0.85), फूल:flower (0.63), 花:flower (0.55)

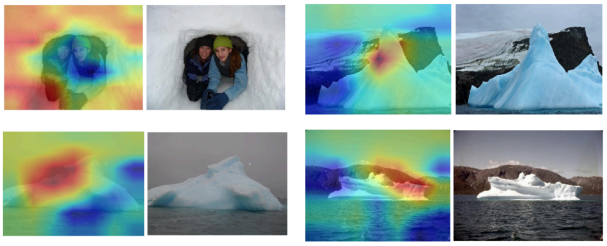


bridge (0.58), पुल:the bridge (0.33), 橋:bridge (0.49)

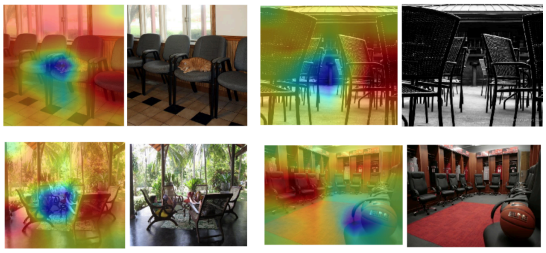


train (0.60), रेलगाड़ी:the train (0.14), 電車:Electric train (0.25)

बैठे:sitting (0.48), 座つ:Sitting (0.61)



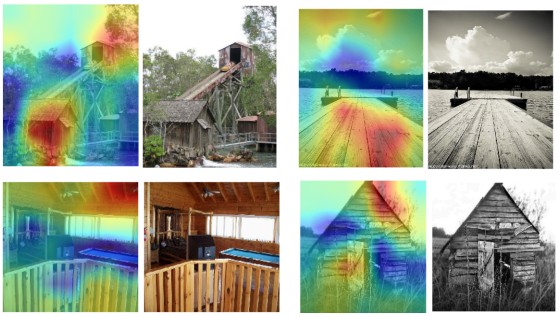
snow (0.70), बर्फ:ice (0.52), 雪:snow (0.41)



chair(s) (0.85), कुर्सियां:chairs (0.63) , 椅子:Chair (0.53)

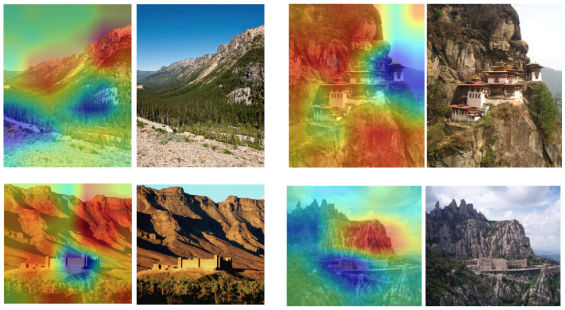
woman (0.74), औरत:the woman (0.44), 女性:Woman (0.62)

yellow (0.8), पीले:yellow (0.61) , 黄色い:yellow (0.55)

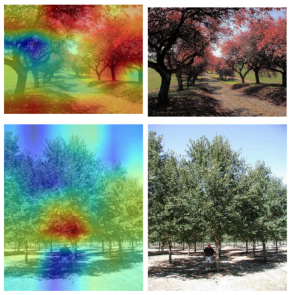


wooden (0.69), लकड़ी:the wood (0.89), 木:wood (0.54)

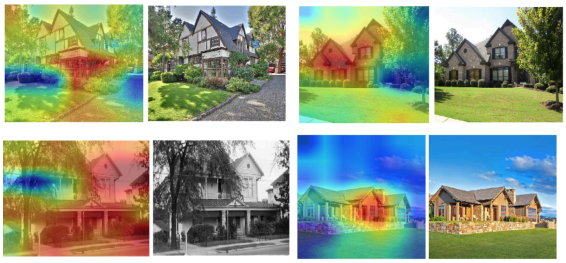
wall(s) (0.78), दीवार:wall (0.61), 壁:wall (0.42)



mountain(s) (0.45), पहाड़:the mountain (0.30), 山:Mountain (0.33)

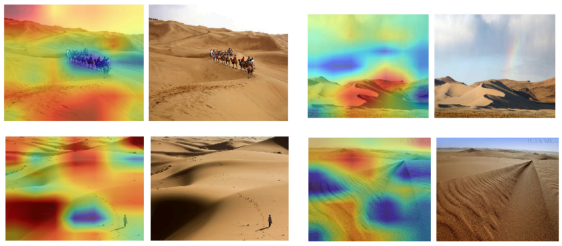


tree(s) (0.79), पेड़:the trees (0.59), 木:wood (0.46)



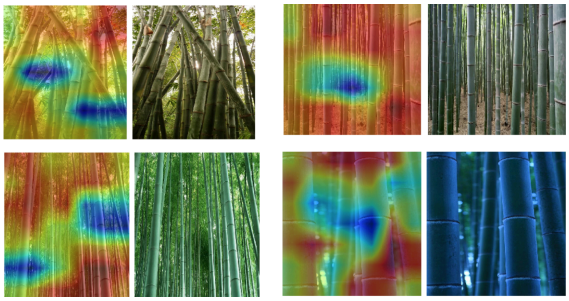
house(s) (0.77), मकान:house (0.40), 家:House (0.31)

shoe(s) (0.79), 靴:shoes (0.53)



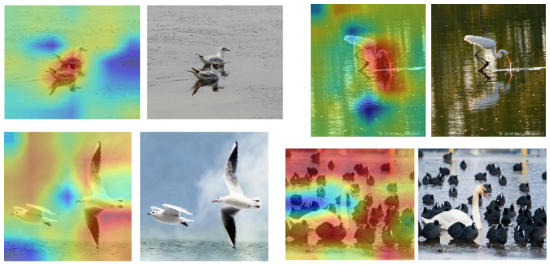
desert (0.34), 砂漠:Desert (0.23)

bowling (0.89), ボーリング:Bowling (0.80)

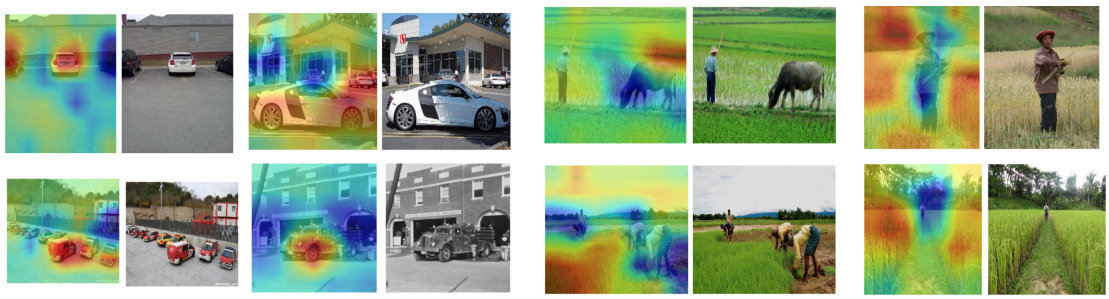


bamboo (0.28), 竹林:Bamboo forest 0.29

flag(s) (0.49), झंडा:flag (0.57), 国旗:Flag (0.42)



bird(s) (0.88), 鳥:bird (0.58)



car(s) (0.53), गाड़ी:the cart (0.71), grass (0.35), घास:grass (0.35), 芝生:lawn (0.40) 車:car(0.43)

Bibliography

[1] NTT corporation, personal communication. 2018.

[2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2017.

[3] Afra Alishahi, Marie Barking, and Grzegorz Chrupała. Encoding of phonology in a recurrent neural model of grounded speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 368–378, 2017.

[4] 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.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *Journal of machine learning research*, 3(Feb):1107–1135, 2003.

[7] Shane Bergsma and Benjamin VanDurme. Learning bilingual lexicons using the visual similarity of labeled web images. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1764–1769, 2011.

[8] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[9] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[10] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 613–622, 2017.

[11] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.

[12] Mark Dredze, Aren Jansen, Glen Coppersmith, and Ken Church. NLP on spoken documents without ASR. In *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, pages 460–470, 2010.

[13] Jennifer Drexler and James Glass. Analysis of audio-visual features for unsupervised speech recognition. In *Proc. GLU 2017 International Workshop on Grounding Language Understanding*, pages 57–61, 2017.

[14] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

[15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[16] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845, 2017.

[17] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[18] David Harwath, Galen Chuang, and James Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4969–4973, 2018.

[19] David Harwath and James Glass. Learning word-like units from joint audio-visual analysis. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 506–517, 2017.

[20] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 649–665, 2018.

[21] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. *International Journal of Computer Vision*, Aug 2019.

[22] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. pages 1858–1866, 2016.

[23] David F Harwath, Timothy J Hazen, and James R Glass. Zero resource spoken audio corpus analysis. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8555–8559, 2013.

[24] William Havard, Jean-Pierre Chevrot, and Laurent 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)*, pages 8618–8622, 2019.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] Aren Jansen, Kenneth Church, and Hynek Hermansky. Towards spoken term discovery at scale with zero resources. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, pages 1676–1679, 2010.

[27] Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*, 2019.

[28] Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater. Unsupervised neural network based feature extraction using weak top-down constraints. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5818–5822, 2015.

[29] Herman Kamper, Aren Jansen, and Sharon Goldwater. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE Transactions on Audio, Speech and Language Processing*, 24(4):669–679, 2016.

[30] Herman Kamper and Michael Roth. Visually grounded cross-lingual keyword spotting in speech. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 248–252, 2018.

[31] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[32] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 48–54, 2003.

[33] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3558–3565, 2014.

[34] Chia-ying Lee and James Glass. A nonparametric bayesian approach to acoustic model discovery. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 40–49, 2012.

[35] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2657–2664, 2014.

[36] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1435–1442, 2012.

[37] Scott Novotney and Chris Callison-Burch. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215, 2010.

[38] Lucas Ondel, Lukáš Burget, and Jan Černockỳ. Variational inference for acoustic unit discovery. *Procedia Computer Science*, 81:80–86, 2016.

[39] Alex Park and James Glass. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech and Language Processing*, 16(1):186–197, 2008.

[40] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069, 2016.

[41] Daniel Renshaw, Herman Kamper, Aren Jansen, and Sharon 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)*, pages 3199ˉ–3203, 2015.

[42] Deb Roy and Alex Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26:113–146, 2002.

[43] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018.

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

[45] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 966–973, 2010.

[46] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond 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*, pages 543–553, 2016.

[47] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[48] Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly translate foreign speech. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, pages 2625–2629, 2017.

[49] Yaodong Zhang, Ruslan Salakhutdinov, Hung-An Chang, and James Glass. Resource configurable spoken query detection using deep boltzmann machines. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5161–5164, 2012.

[50] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

[51] Bowen Zhou. Statistical machine translation for speech: A perspective on structures, learning, and decoding. *Proceedings of the IEEE Special Issue on Speech Information Processing*, 101(5):1180–1202, 2013.