SECOND LANGUAGE TRANSFER LEARNING IN HUMANS AND MACHINES USING IMAGE SUPERVISION

Kiran Praveen[1*], *Anshul Gupta*[1,2*], *Akshara Soman*[1], *Sriram Ganapathy*[1]

[1]Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore.
[2] International Institute of Information Technology, Bangalore.

ABSTRACT

In the task of language learning, humans exhibit remarkable ability to learn new words from a foreign language with very few instances of image supervision. The question therefore is whether such transfer learning efficiency can be simulated in machines. In this paper, we propose a deep semantic model for transfer learning words from a foreign language (Japanese) using image supervision. The proposed model is a deep audio-visual correspondence network that uses a proxy based triplet loss. The model is trained with large dataset of multi-modal speech/image input in the native language (English). Then, a subset of the model parameters of the audio network are transfer learned to the foreign language words using proxy vectors from the image modality. Using the proxy based learning approach, we show that the proposed machine model achieves transfer learning performance for an image retrieval task which is comparable to the human performance. We also present an analysis that contrasts the errors made by humans and machines in this task.

Index Terms— Multimodal learning, transfer learning, document retrieval, human-machine comparison, distance metric learning

1. INTRODUCTION

The early work done by Locke [1] on language understanding suggests that "To make children understand what the names of simple ideas or substances stand for, people ordinarily show them the thing whereof they would have them have the idea and then repeat to them the name that stands for it." The acquisition of words from a second language in both children and adults exhibit remarkable similarity [2]. The term 'second language' (also called L2) refers to any language that is not one's native language (also called L1) [3]. There is evidence to show that translation equivalents are linked at a conceptual level in bilinguals [4]. Further, other studies have shown that a person's L1 influence his L2 acquisition. For instance, [5] showed that cognates, or words with a common etymological origin, are easier to learn. While there are several methods for novel word learning, Pavlenko [6] noted that the picture-naming task is the only task that taps into the mapping between words and their real-world referents. Studies have shown that semantic knowledge is grounded in the perceptual space [7], and concrete nouns with their richer morphological representations are better learned with image supervision [5]. It is also shown that humans require only a very small number of instances to learn meanings of new words [8]. It is therefore of significant interest to question whether machines can achieve such efficiency with limited data.

In the recent years, advances in deep learning methods have enabled machines achieve human-like performance on several tasks like speech recognition [9], machine translation [10], computer vision [11] and face detection [12]. In many of these paradigms, the deep models use significantly large amounts of data compared to the humans which makes them vulnerable in limited data scenarios like transfer-learning and adaptation tasks. In this paper, we propose a deep semantic model that can achieve human-like performance for a transfer learning task. To the best of our knowledge, this paper is the first attempt to compare humans and machines for a language transfer learning task.

We explore a rapid language learning task where human subjects attempt to learn a set of words from a new language with image supervision. The subjects are provided with only two instances for each novel word along with the corresponding image. The human subjects in our experiment show a high semantic recall accuracy. Given this rapid learning ability achieved by humans, we investigate whether modern deep networks with transfer learning methods can emulate this task. Specifically, we propose a deep semantic model to mimic this transfer learning task that uses a proxy based learning mechanism [13] on multi-modal (audio/image) inputs. The model, trained on large dataset of familiar language (English) words and their corresponding image counterparts, is transfer learned with small number of examples from a foreign language (Japanese). The performance of the proposed model is compared with the human performance on cross-modal image retrieval where we show that the proposed approach can achieve human-like performance on this transfer learning task. In a subsequent error analysis, we also highlight that the errors made by the deep model are quite different compared to those made by the human subjects.

2. RELATED PRIOR WORK

Multi-modal modeling of image and text has received significant attention in the recent years. Barnard et al. in an early work [14] relied on labeled images to estimate the joint distribution between words and objects. The work by Socher [15] learned a latent semantic space covering images and words learned on completely non-parallel data. In the recent past, the success of recurrent deep neural networks [16, 17] have generated much interest in the field of visual-text modeling. For modeling the joint semantic space of audio and images, a deep neural network model capable of spoken language acquisition from untranscribed audio training data was presented in [18] where the only supervision comes from contextually relevant visual images. Here, the authors use spoken audio captions for an image dataset and the model is evaluated for an image annotation

[*] denotes equal contribution

This work was funded by grants from the Pratiksha Trust and DST Early Career Award (ECR 2017/1341).

978-1-7281-0306-8/19/$31.00 ©2019 IEEE

Block	Words
1	hachinosu(*beehive*), doramu(*drum*), kouhan(*plates*), shippu(*ship*), torakku(*truck*)
2	ari(*ant*), kyoukai(*church*), sakana(*fish*), miruku(*milk*), hanarabi(*teeth*)
3	mokkori(*bedsheets*), busu(*booth*), kaji(*fire*), yubiwa(*ring*), sukkurin(*screen*)
4	kaeru(*frogs*), mendori(*hen*), mune(*lungs*), puru(*pool*), tento(*tent*)
5	shuzu(*shoes*), taiyou(*sun*), takushi(*taxi*), cha(*tea*), shita(*tongue*)
6	keki(*cake*), isu(*chair*), hanabana(*flowers*), maggu(*mugs*), komugi(*wheat*)
7	ginkou(*bank*), nesuto(*nest*), yakuzai(*tablets*), dorobou(*thief*), torappu(*trap*)
8	koushi(*calf*), kyappu(*cap*), doa(*door*), matto(*mat*), shio(*tide*)
9	houki(*broom*), koin(*coin*), nedoko(*cot*), resutoran(*restaurant*), supun(*spoon*)
10	hikouki(*aeroplane*), kaban(*bag*), omeme(*eyes*), kaito(*kite*), sokkusu(*socks*)
11	tokei(*clock*), kuran(*crown*), genkotsu(*fist*), nobu(*knob*), suwan(*swan*)
12	benchi(*bench*), ushi(*cow*), purezento(*gifts*), kagi(*keys*), jumoku(*tree*)
13	bouru(*bowl*), nezumi(*mouse*), koromo(*robe*), ropu(*rope*), nagashi(*sink*)
14	hako(*box*), naifu(*knife*), jou(*lock*), toge(*thorns*), besuto(*vest*)
15	nedoko(*cot*), kouzui(*flood*), medaru(*medal*), moppu(*mop*), hoshiboshi(*stars*)
16	pasupoto(*passport*), pai(*pie*), satsu(*police*), sekken(*soap*), kitte(*stamp*)
17	kamera(*camera*), gaun(*gown*), zubon(*pants*), hitsuji(*sheep*), ryourin(*wheels*)
18	hon(*books*), jusu(*juice*), mappu(*map*), kasha(*van*), rou(*wax*)

Table 1. Words used for human learning task divided into experimental blocks and colored into *Hiragana* and *Katakana* words along with the English translation.

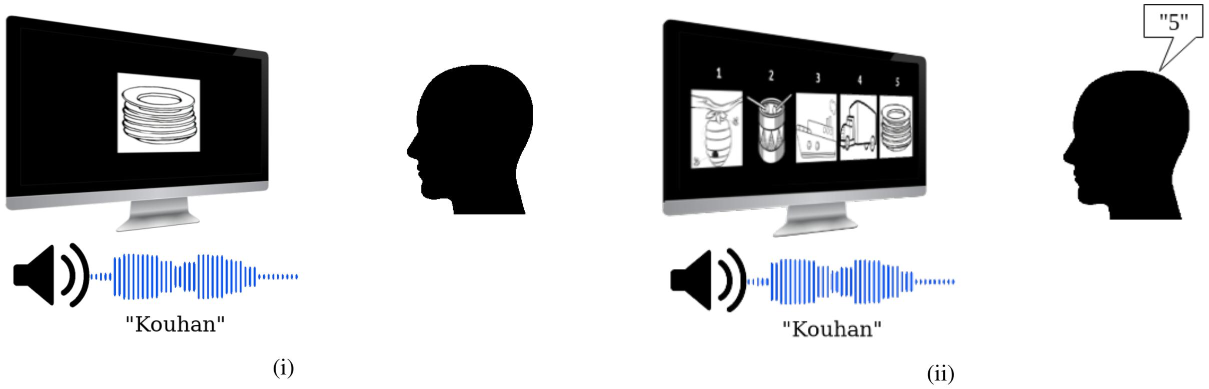


Fig. 1. (i) Image supervision for the word /kouhan/ in the learning phase, and (ii) recall question in the testing phase for /kouhan/. Note that all five words of the block are introduced in the learning phase before recall.

task. In a similar manner, the design of a system to learn both audio and visual semantic information in a audio-visual correspondence task was attempted in [19]. However, the work is directed toward natural scenes and environmental audio recordings.

The concept of transfer learning has been well explored in computer vision. In [20], the authors trained an image classifier on the ImageNet database [21] and were successfully able to transfer representations to the Pascal VOC database [22]. Similarly, [23] used representations learned from training on the ImageNet database to achieve state of the art performance on the Caltech-101 [24] and Caltech-256 [25] datasets. Transfer learning for speech and audio has been investigated in [26] to enhance noisy mandarin speech data using a DNN model trained only on English and vice versa. In a recent work [27], the authors developed deep models for transfer learning from environmental sound classification to a speech task.

In this paper, we propose a deep semantic model which is first trained using multi-modal image/speech (English) input. The model attempts to learn the semantic correspondence between speech and image inputs. We use a novel modification to the proxy based approach proposed recently in [13]. Then, the representations learned are transferred for Japanese language speech input. We show that the proposed deep semantic model achieves human-like performance for the transfer learning task.

3. HUMAN EXPERIMENT

The participants were Indian nationals with self-reported normal hearing and no history of neurological disorders. Twelve adults participated in this study (mean age = 24.5, age span = 22-28, 6 female and 6 male) who had an intermediate or higher level of English proficiency. It was verified with the Oxford Listening Level Test [28] before the commencement of the experiment. These subjects had no prior exposure to Japanese language. In our human behavioral experiment, we choose Japanese as the novel language as it does not belong to the Indo-European language family. There is very little similarity with English and native Indian languages. At the same time, Japanese contains a set of loan-words from English termed as *Katakana* words. Katakana words are typically English words that have been adapted without translation into the Japanese language [29]. These sound very similar to their English counterparts. By us-

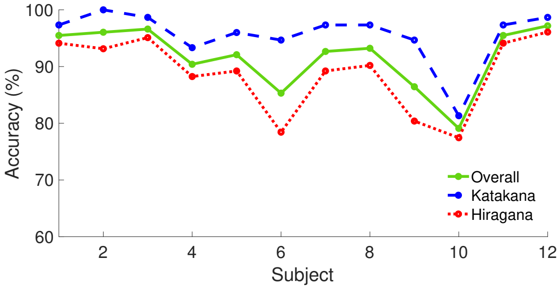


Fig. 2. Human behavioral experiment:subject-wise accuracy for semantic retrieval task.

ing a mix of native Japanese words (referred to as Hiragana words) and Katakana words, we can test for the effect of phonological similarity in learning. The set of words used in our experiments (Table 1) consists of 38 Katakana words and 52 Hiragana words. The word stimuli used are human voices spoken by a fluent Japanese speaker.

3.1. Learning Phase

In the learning phase, the subjects heard one word at a time and the corresponding image. The images serve as an anchor that helps map the Japanese word to a common semantic representation of English. This process is repeated for five words to form a block. A screenshot of the learning phase is shown in Fig 1 (i). Every word image pair is only provided twice.

3.2. Testing Phase

In the testing phase, the subjects heard a word from the block of 5 words. They are then asked to match it against all the five images from that block. This is effectively the human equivalent of the image retrieval task described later for the machine model. We perform audio queries for each word in the block in a different order to the one used in the learning. We report the retrieval accuracy of each subject averaged over all 18 blocks. A screenshot of the image used in the testing phase is shown in Fig 1 (ii).

3.3. Results

The results for the human experiment are shown in Fig 2. The average human accuracy is 91.7% while there is on the average 7% higher accuracy for Katakana words compared to Hiragana words. The results show that humans are remarkably efficient in learning new words semantically with image supervision.

4. DEEP SEMANTIC MODEL

4.1. Dataset

The first challenge towards the semantic modeling of audio-image modalities is the creation of a suitable dataset. Since we could not find an open source parallel corpus of object images and their corresponding audio recordings, we generated a new dataset for this task using images from image datasets and audio from synthesized speech of the labels. As the modern text to speech synthesis systems have reported human quality speech outputs especially for short duration speech generation tasks like words [30], we use the synthesized audio for most of the experiments in this work (we also report testing the model with human recorded Japanese audio).

Table 2. Details of the dataset. The numbers are listed for each class with a total of 655 classes used in the final model.

Data	Train	Validation	Test
Image (ImageNet classes)	160 ImageNet train	16 ImageNet val	16 ImageNet val
Image (New classes)	60 30 Google 30 Flickr	20 10 Google 10 Flickr	20 10 Google 10 Flickr
Speech (TTS voice) (English)	22 10 Google 1 IBM 11 Microsoft	5 2 Google 1 IBM 2 Microsoft	5 2 Google 1 IBM 2 Microsoft
Speech (TTS voice) (Japanese)	3 1 Google 1 IBM 1 Microsoft	1 1 Microsoft	1 1 Microsoft

For the images, we have used a subset of the ImageNet [21] database by selecting 576 classes and added additional images from Flickr and Google image repository giving a total of 655 classes. The 90 objects used in the human experiment (Table 1) are part of the 655 classes. All the audio recordings for each image class is of one word length. The ImageNet dataset contains 1000 images per class for training and 50 images per class for validation. The audio recordings are generated using Google [31], IBM [32], and Microsoft [33] Text-to-Speech (TTS) systems. More details on train, validation and test data used in our experiments are given in Table 2.

4.2. Audio-Visual Semantic Network

The audio-visual semantic network used in this work is illustrated in Fig 3. The model has audio and image sub-networks which are trained jointly using the multi-modal input.

4.2.1. Audio sub-network

The audio sub-network consists of two recurrent layers with LSTM units followed by fully connected layers. We use a train, validation and test split as given in Table 2. The audio sub-network is initialized by training in a classifier setting where the task is to classify among the 655 classes. The training data is augmented with 6 different types of noise to increase the variability with 88704 samples for training, and 2880 samples for validation and test respectively. We use 80 dimensional bottleneck features (BNF) from a deep neural network (DNN) trained for automatic speech recognition (ASR) on the switch-board and Fisher corpora [34]. In the pre-training task, the BNF features give superior performance for a top 1 classification accuracy of 84.3% (compared to 59.1 % obtained for conventional mel-frequency features). We use the 2048 dimensional pre-softmax layer as an embedding to represent the audio recording. The audio sub-network weights (of network S1 and S2 in Fig 3) are learned for the English audio-visual correspondence task. The sub-network weights for S2 are transferred to Japanese speech.

4.2.2. Image sub-network

The image sub-network is the Xception network [35] which is pre-trained on the ImageNet database. The Xception network is essentially an extension of the Inception architecture [36] and replaces

Code available at https://github.com/Anshul-Gupta24/Audio-Visual-Deep-Multimodal-Networks

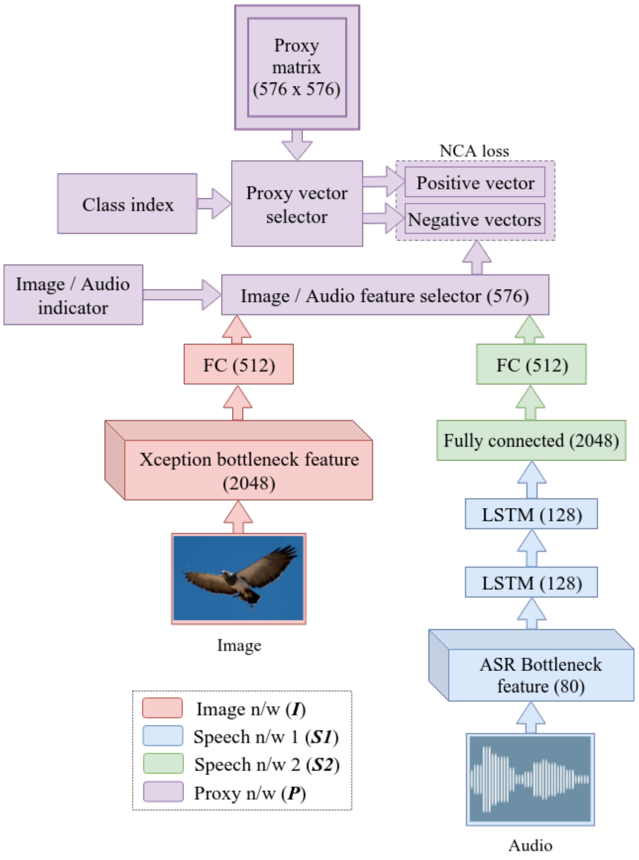


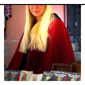
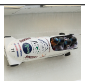
Fig. 3. Joint audio-visual semantic network trained with proxy NCA loss. The audio network consisting of S1 and S2 are learned for English while S2 alone is transfer learned for Japanese.

the Inception modules with modified depth wise separable convolution layers to entirely decouple the mapping of cross channel correlations and spatial correlations in the feature maps. The modified depth wise separable convolution operation performs 1×1 convolution followed by spatial convolution over each of the output channels to reduce the number of operations. We use the 2048 dimensional pre-softmax layer as an embedding to represent our image. The image sub-network consists of the 2048 dimensional embedding layer followed by a fully connected layer. These weights are re-trained for the audio-visual correspondence task.

Table 3. Accuracy for image retrieval with English audio task. Chance accuracy is 0.15%

Model	**Image retrieval(%)**		
	Top-1	**Top-5**	**Top-10**
Entropy	23.70	42.80	51.10
Triplet	47.20	78.20	84.80
Proxy	62.60	77.30	81.00

Table 4. Examples of Top-3 image retrieval outputs for audio query in English. In some cases, the common audio syllables are evident in the top confusions (example /bobsled/), while other cases it is the visual similarity that causes the confusions (example /centipede/)

Speech query	**Image retrieval result**		
	1	**2**	**3**
cassette	 cloak	 *cassette*	 potpie
bobsled	 *bobsled*	 snowmobile	 limousine
centipede	 isopod	 flatworm	 ant

4.3. Joint Audio-Image Training

We train the audio-visual correspondence model to learn the joint semantic distribution of the image inputs and English audio inputs. The model is learned in such a way that the similarity between matching audio-image pairs is high, while those between non-matching audio-image pairs is low. Traditionally, for this problem, supervision is expressed in the form of triplets. A main issue is the need for finding informative triplets by tricks like hard or semi-hard triplet mining. Even with these tricks, the convergence rate of such methods is quite slow and does not generalize to transfer learning tasks [13]. In this paper, we optimize the triplet loss on a different space of triplets, consisting of an anchor data point and similar and dissimilar proxy points which are learned as well. We modify the original proxy approach [13] by using a proxy based triplet loss that maximizes similarity between an anchor data point and the matching proxy, while minimizing similarity with the non-matching proxies. This way the model can learn from multi-modal data.

We use cosine similarity as a similarity measure and minimize the neighborhood component analysis (NCA) loss [37]. The similarity measure is given by,

$$S_{k,l} = \{1_{\text{anchor=image}}\} y_k^T p_l + \{1_{\text{anchor=audio}}\} x_k^T p_l \qquad (1)$$

where y_k, x_k and p_l are the L2 normalized embeddings for image k, audio k and proxy l respectively. The NCA loss is given by,

$$C_{k,l}(\theta) = -\log \frac{\exp(S_{k,l})}{\sum_{p \in L, p \neq l} \exp(S_{k,p})} \qquad (2)$$

where θ are the model parameters, k is an input anchor and l is its corresponding proxy, while L is the set of non-corresponding proxies. The proxy vectors p_l are real vectors which are unique for each class. Since we have the class information for every data sample, both the weight parameters of the model as well as the proxy vectors can be jointly learned [13]. We train for about 100 epochs for the audio-image correspondence task with a learning rate of $1e^{-2}$. We use the Adam [38] optimizer with batch-norm [39].

The advantage of such a two step process is that the proxy vectors do not need to be learned again as the images remain constant across the languages. Hence, during the language transfer-learning, we skip the proxy learning process and train only the weights of the audio sub-network.

4.3.1. *Results on Image Retrieval Task*

On the English audio-image semantic correspondence task, we test the image retrieval accuracy by providing an audio query and finding which test image matches the audio the most (among random test

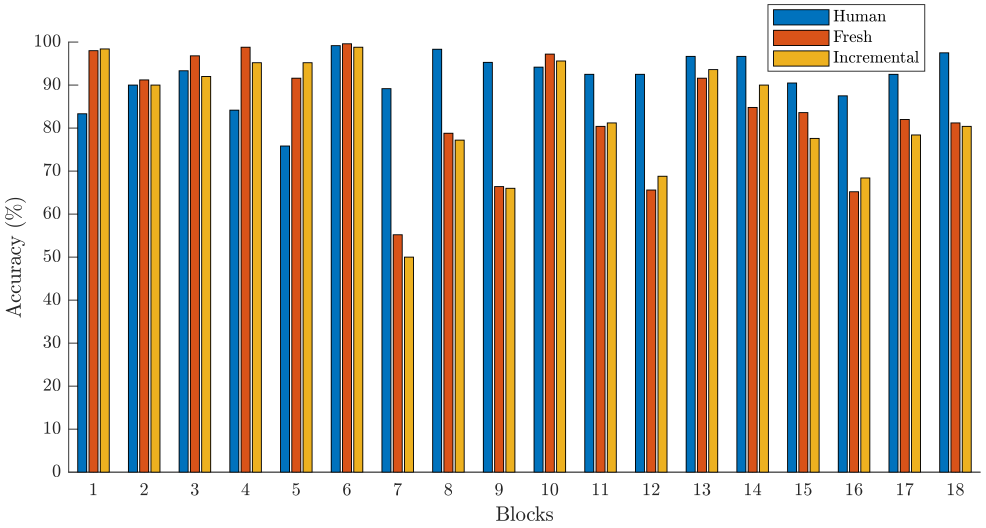


Fig. 4. Retrieval accuracy of all blocks in order of their appearance in human experiment

images from all classes). If the class of image that has the best similarity score with the audio input also matches with semantic class of the audio, then it is counted as a hit in the accuracy measure. In Table 3, we report the average image retrieval accuracy (20 queries per class) for different choice of loss functions in the model configuration. While the top-5 and top-10 accuracies are better for the triplet model, using a binary cross-entropy (entropy) or the standard triplet loss gives a performance for the image retrieval task (top-1) that is worse than the proposed proxy based approach. Hence, we use the proxy loss model for the transfer learning task. Typical examples of confusions from the image retrieval task are shown in Table 4. It is interesting to note that the model confusions are not totally arbitrary and this could be due to acoustic similarity (example /bobsled/) or image similarity (example /centipede/).

4.4. Transfer Learning

The audio sub-network in the deep semantic model is transfer learned for Japanese speech inputs. This subset is the same set of words presented to the human subjects. In order to emulate the human experiment, we consider 5 audio classes at a time corresponding to each block shown to the subjects (Table 1). The fully connected layer in the audio sub-network is re-trained in the transfer learning process (as the amount of transfer learning data in each block is small). The number of epochs for transfer learning is a hyper parameter as the model can overfit to the small number of examples. The transfer learning is done in two ways,

- **Fresh start** - at the start of every block, the model weights are initialised with the weights from the English model.
- **Incremental start** - the chronological order of the blocks appearing in the human experiment is considered during training. The initial model in the current block is the final trained output model from the previous block.

The vision sub-network of the proxy model along with the proxy matrix remain fixed during training. The weights of the fully connected layers in audio sub-network after LSTM are kept trainable. Each block has 15 speech samples for train, 5 speech samples for validation and testing respectively. The details of the dataset used in the transfer learning are given in Table 2.

Table 5. Performance comparison of human and machine model for the transfer learning task. Chance accuracy is 20%

Model	Human	Fresh	Incremental
Overall	91.67	83.78	83.16
Hiragana	88.81	85.00	83.62
Katakana	95.56	82.11	82.53

5. RESULTS AND DISCUSSION

5.1. Machine vs. human performance

Figure 4 shows the block-wise performance of humans and machines. The model achieves an accuracy which is comparable to humans. The human performance is much more stable across the blocks while the machine models have more variance in the results. Overall, the human performance is better than the machine models for 11 of the 18 blocks considered. However, it is interesting to note that human models get better than machine models for the last part (humans are better than machine models for 11 of the last 11 blocks). This plot shows that humans are incrementally better at learning (given the same semantic complexity of learning) compared to machines. The incremental learning in the machine model did not show particular improvements compared to the fresh start learning. Using unsynthesised audio data for testing drops the average performance only by approximately 8%.

5.2. Hiragana versus Katakana accuracy

Figure 7 shows the retrieval accuracy of Katakana words compared to Hiragana words. Initially there is a significant advantage for Katakana words compared to Hiragana words, but the model does not show any bias for Katakana words after a few epochs. As seen in Table 5, humans show better recall accuracy with Katakana words while the machine models don't show a significant change in accuracy for Katakana over Hiragana words. This highlights that humans may be better able to parse word strings to sub-word units and are more resilient to modifications of the sub-word unit acoustics.

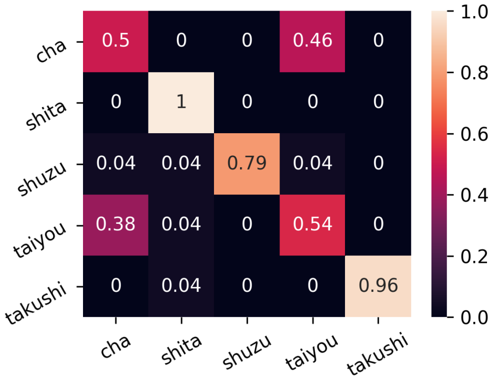


Fig. 5. Confusion matrix - hardest block in human exp. (Block 5).

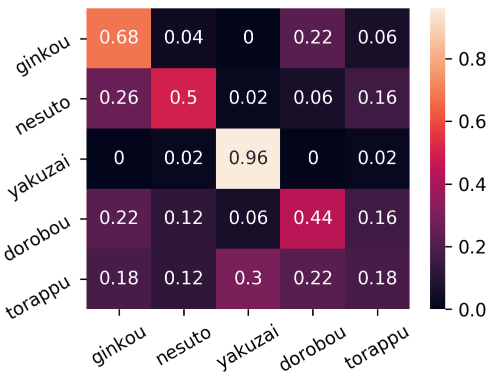


Fig. 6. Confusion matrix - hardest block in machine model (Block 7).

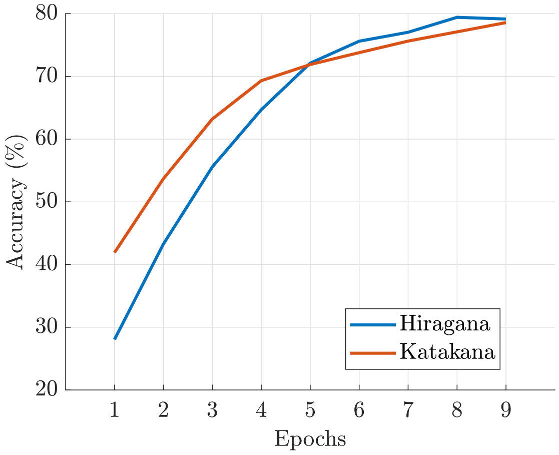


Fig. 7. Transfer learning accuracy for Katakana and Hiragana words.

Table 6. Accuracy of the model on Japanese audio image retrieval on 90 classes. Chance accuracy is 1.11%

	Accuracy
Transfer	35.04%
No transfer	1.09%

5.3. Results Without English Initialization

To check whether transfer learning is useful, the speech section of the proxy network is randomly initialised and trained for image retrieval on all 90 classes in the experiment (one single block of 90 without separate blocks of 5 done in previous experiments). The results shown in table 6 indicates that English learning provides significant gains in learning the image-audio correspondence for the Japanese words.

5.4. Hard Examples for Humans and Machines

The confusion matrix for blocks with the worst performance for humans and machines are shown as a confusion diagram in figures 5 and 6 respectively. The y-axis of the figures correspond to the speech query and x-axis correspond to the image retrieved. There is a symmetry which can be observed in the error patterns in humans, whereas the confusion matrix of machine model does not illustrate symmetry. The confusion matrix in the machine model is also more dense indicating confusions among more output classes for the examples from the given target class.

Table 7. Accuracy of image retrieval (using English audio) for 90 classes used in the human experiment with different playback speeds. Chance accuracy is 1.11%

Speed	**0.8x**	**1.2x**	**1x**
Accuracy (%)	38.51	16.09	49.07

5.5. Perturbation analysis

To further illustrate the vulnerability of the model in learning the Katakana words, we investigate the robustness of the model to time scale modifications of the audio stimuli. The results in Table 7 show that when the playback is slowed down or sped up, the retrieval performance is significantly worse. In informal listening tests, humans did not find significant change in performance for speed perturbations. This may explain the difference in model performance for Katakana words between the human and machine models.

6. CONCLUSION

The following are the major contributions from the work,

- Design a novel paradigm to analyze human performance in language transfer learning for cross-modal image retrieval.
- Develop a deep semantic model for learning the audio-image correspondence in image retrieval task. The proxy based loss improves the retrieval accuracy significantly.
- Analyze the transfer learning performance of the deep semantic model and compare the performance with human results.

The difference in Katakana performance highlights that humans may be better able to parse word strings to sub-word units and are more resilient to modifications of the sub-word unit acoustics. In the future we plan to investigate this hypothesis using methods such as including speed perturbed audio in training.

7. ACKNOWLEDGEMENTS

We would like to acknowledge Axxonet Systems and Dr.Arun Sasidharan in paritcular for his help in data collection, and Pratibha R. for the experimental protocol, data analysis and collection.

8. REFERENCES

[1] John Locke, "An essay concerning human understanding (pp. 259–298)," *Cleveland, Ohio: Meridian Books.(Original work published in 1690)*, 1964.

[2] David M Singleton, *Exploring the second language mental lexicon*, Ernst Klett Sprachen, 1999.

[3] "Second language," https://dictionary.cambridge.org/dic\tionary/english/second-language, Last accessed June 22, 2019.

[4] Jeanette Altarriba, "The representation of translation equivalents in bilingual memory," *Cognitive Processing in Bilinguals*, vol. 83, 12 1992.

[5] Annette MB De Groot and Rineke Keijzer, "What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting," *Language learning*, vol. 50, no. 1, pp. 1–56, 2000.

[6] Aneta Pavlenko, "Conceptual representation in the bilingual lexicon and second language vocabulary learning," *The bilingual mental lexicon: Interdisciplinary approaches*, pp. 125–160, 2009.

[7] Lawrence W Barsalou, "Grounded cognition," *Annual Review of Psychology*, vol. 59, pp. 617–645, 2008.

[8] Arielle Borovsky, Jeffrey L Elman, and Marta Kutas, "Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context," *Language Learning and Development*, vol. 8, no. 3, pp. 278–302, 2012.

[9] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 12, pp. 2410–2423, 2017.

[10] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al., "Achieving human parity on automatic chinese to english news translation," *arXiv preprint arXiv:1803.05567*, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[12] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[13] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh, "No fuss distance metric learning using proxies," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 360–368.

[14] 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*, vol. 3, no. Feb, pp. 1107–1135, 2003.

[15] 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*. IEEE, 2010, pp. 966–973.

[16] 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*, 2015, pp. 3128–3137.

[17] 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*, 2015, pp. 3156–3164.

[18] David Harwath, Antonio Torralba, and James Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing Systems*, 2016, pp. 1858–1866.

[19] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[20] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012)," http://host.robots.ox.ac.uk/pascal/VOC/voc2012/, Last accessed June 22, 2019.

[23] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[24] Li Fei-Fei, Rob Fergus, and Pietro Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer vision and Image understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[25] Gregory Griffin, Alex Holub, and Pietro Perona, "Caltech-256 object category dataset," 2007.

[26] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Cross-language transfer learning for deep neural network based speech enhancement," in *The 9th International Symposium on Chinese Spoken Language Processing*. IEEE, 2014, pp. 336–340.

[27] Brian McMahan and Delip Rao, "Listening to the world improves speech command recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] "Oxford listening level test, https://www.oxfordonlineenglish.com/english-level-test/listening," .

[29] Ben Olah, "English loanwords in japanese: Effects, attitudes and usage as a means of improving spoken english ability," *Bunkyo Gakuin Daigaku Ningen Gakubu Kenkyū Kiyo*, vol. 9, no. 1, pp. 177–188, 2007.

[30] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio.," in *Proceedings of the 9th Speech Synthesis Workshop*, 2016.

[31] "Google cloud text-to-speech," https://cloud.google.com/text-to-speech/, Last accessed June 22, 2019.

[32] "Ibm watson text-to-speech," https://www.ibm.com/watson/services/text-to-speech/, Last accessed June 22, 2019.

[33] "Microsoft azure text-to-speech," https://azure.microsoft.com/en-in/services/cognitive-services/text-to-speech/, Last accessed June 22, 2019.

[34] Seyed Omid Sadjadi, Timothée Kheyrkhah, Audrey Tong, Craig Greenberg, Douglas Reynolds, Elliot Singer, Lisa Mason, and Jaime Hernandez-Cordero, "The 2017 nist language recognition evaluation," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 82–89.

[35] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[37] Jacob Goldberger, Geoffrey E Hinton, Sam T Roweis, and Ruslan R Salakhutdinov, "Neighbourhood components analysis," in *Advances in neural information processing systems*, 2005, pp. 513–520.

[38] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.