6.867 Machine Learning Final Project December 13th, 2016

Learning Digits via Audio-Visual Representations

Prof. Kaebling, Sra, Lozano-Perez *Andrew Xia, Karan Kashyap, Sitara Persad*

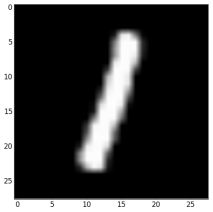
1 Abstract

Our goal is to explore models for language learning (in this case learning numerical digits in their spoken and visual representations) in the manner that humans learn languages as children. Namely, children do not have intermediary text transcriptions in corresponding visual and audio inputs from the world around them; rather, they directly make connections between what they see and what they hear. In this paper, we construct models for the direct bi-directional classification of speech and images, inspired by a few research papers: [1] [2]. We experiment with architectures of two convolutional neural networks, one on the TIDIGITS data set (audio) and the other on the MNIST data set (visual), to obtain joint representations of single digits from spoken utterances and images. Finally, we experiment with an alignment model that ties together the convnets to learn these joint representations. We report an overall image annotation accuracy of 88.5% and an overall image retrieval accuracy of 87.6%.

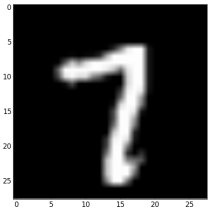
2 Introduction and Related Work

Typically, training good automatic speech recognition models is a data-intensive endeavor and involves highly supervised learning where audio inputs are paired with text transcriptions. Instead of textual annotations, we explore matchings of audio data directly, in the form of spoken utterances of numerical digits, with corresponding images of the digits taken from the MNIST data set. Bypassing the text component in the task of learning correspondences between images and speech, especially in larger real-world sets of data, may see performance gains compared to methods that use a text-based intermediary approach.

In this paper, we use convolutional neural networks to derive understanding of image objects directly from speech, and vice versa, as previously explored by Harwath et al. [1] [2]. As mentioned, this problem is usually explored with accompanying text transcriptions, which act as an



(a) MNIST 1



(b) MNIST 7

Figure 1: Examples of MNIST hand-drawn digits 1 and 7.

intermediary step between audio and visual data. We experiment with a pair of convolutional neural networks for both speech spectrogram and image inputs and tie the convnets together using an embedding and alignment model as described in [1].

Our image data set consists of 28×28-pixel hand-drawn digits from the MNIST corpus, which contains 60,000 training samples and 5,000 test samples. Our audio data set consists of training and test sets of 4,891 and 2,325 audio files of single digit utterances, respectively. This was a subset of the recordings from the TIDIGITS data set (only adults; excluding children), as other recordings in the TIDIGITS dataset included mutiple digit utterances. We convert each spoken utterance into a spectrogram with 23 frequency bins and over 100 frames totaling 1 second.

This approach closely follows previous research by David Harwath and James Glass [1]. In their paper, Harwath and Glass present a model which can learn a joint semantic representation over spoken words as well as visual objects using the Flickr8k image set and spoken captions collected from Amazon Mechanical Turk workers. Unlike our spoken samples which are single digit utterances, the captions used by Harwath et al. consist of multiple-word sentences. Thus we reproduce their methods but on the much simpler MNIST and TIDIGITS data sets, limiting the scope to single digits.

The MNIST convnet achieved a validation accuracy of 99.3%. We explored two architectures for the spectrogram convnet: using 512-unit fully connected layers achieved validation accuracy of 89.6%, and 1024 units achieved 95.3%. We used the latter architecture to tune our alignment model using stochastic gradient descent. We define two metrics to evaluate our alignment model. (1) We conduct *image annotation* by taking a MNIST image and matching it with the best spoken utterance. (2) We also conduct *image retrieval* by taking a TIDIGITS utterance and returning the best matching MNIST image. We report an overall image annotation accuracy of 88.5% and an overall image retrieval accuracy of 87.6%.

There are many possible directions we can further take this research. We could experiment with multiple digit representations, in which an image contains multiple MNIST digits and an utterance includes multiple spoken digits. We could also extend our audio samples to languages beyond English, and explore machine translation routes.

3 Methods

We used the MNIST data set of hand-drawn digits for images and the TIDIGITS data set which contains recordings of digit sequences read aloud by humans. Since we wanted to focus on single digit utterances, we filtered the TIDIGITS set and took only the subset of recordings that were single digit utterances. Although TIDIGITS has recordings from both adults and children, we chose to focus on the adult recordings (both men and women) for our task, as we wanted to start with a more homogeneous data set.

3.1 Preprocessing TIDIGITS audio files

We first converted the TIDIGITS files into *.wav* files, as the TIDIGITS data was recorded in 1993 in a deprecated flac compressed wav audio format [3]. We then filtered out the data to include only single digit utterances of adults. We convert each single digit spoken utterance from TIDIGITS into

a log mel filterbank spectrogram using the Kaldi toolkit [4] with 23 frequency bins, with each frame represents a 10 millisecond unit of time. Figure 2 presents visualizations of the log mel filterbank spectrograms of "one" and "seven" utterances from TIDIGITS, generated with the Kaldi toolkit. Note that the number of frames varies based on the length of each recording.

To input into the spectrogram convnet, we want to ensure dimensional uniformity across our spectrograms. We chose to fix the frame size to 100 (1 second long) by zero-padding and truncating our spectrograms as necessary but making sure to center them as specified in [1]. Figure 3 presents a visualization of a post-processed spectrogram for the utterance "oh" (zero) as would be fed into the spectrogram convnet.

3.2 TIDIGITS Spectrogram Convnet

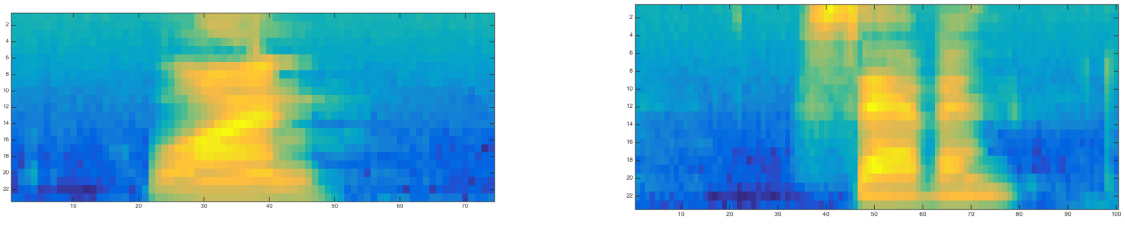
Basing our design off the spectrogram convnet from [1], we derive the following network structure, where we experiment with $N = 512$ and $N = 1024$ for the fully connected hidden units:

1. Convolutional layer with filter size 5×23, stride of 1, top and bottom padding of 1, and 64 output channels with a ReLU nonlinearity;
2. Max pooling layer with height 3, width 4, vertical stride 1, and horizontal stride 2;
3. Two fully connected layers with N hidden units each and ReLU nonlinearities;
4. A softmax classification layer.

Note that we are classifying over eleven classes in this case, as the digit 0 has two spoken representations: *oh* and *zero*. We find that $N = 1024$ performs better, so we use this architecture in the alignment model (Section 3.4).

3.3 MNIST Convnet

We experimented with various architectures for the MNIST convnet, but found the architecture used in a TensorFlow advanced tutorial was optimal. The MNIST convnet trained on 60,000 training samples, first preprocessing the 28×28 MNIST images by normalizing the [0,255] values



(a) Spectrogram of "one"

(b) Spectrogram of "seven"

Figure 2: (a) Spectrogram generated for the utterance "one" with 23 frequency bins over 100 frames (1 second total). (b) Spectrogram for "seven" with specifications similar to (a).

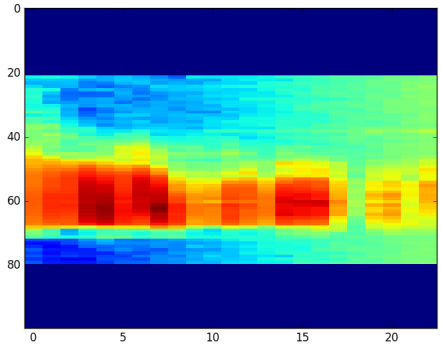


Figure 3: Processed spectrogram for the utterance "oh" for the number zero with zero-padding and centering for 100 frames.

to [-0.5,0.5]. We ran the convnet using batch size of 64 over 10 epochs. The convnet used the following architecture:

1. Convolutional layer with filter size 5, stride of 1, and 32 output channels with a ReLU nonlinearity;
2. Max pooling layer with window size and stride of 2;
3. Second convolutional layer with filter size 5, stride of 1, and 64 output channels with a ReLU nonlinearity;
4. Second max pooling layer with window size and stride of 2;
5. Two fully connected layers with 512 hidden units each and ReLU nonlinearities;
6. A softmax classification layer.

3.4 Alignment Model

Once satisfied with the individual performances of our convnet architectures, we tie our models together to associate the MNIST images with TIDIGITS audio. We base the alignment model off our reference paper [1] but make some variations to the stochastic gradient descent (SGD) objective function and final score calculation. As in [1], in order to obtain vector representations of our TIDIGITS spectrograms and MNIST images we feed each input into the respective convnet and, ignoring the softmax outputs, save the activations (512-dimension for MNIST and 1024-dimension for TIDIGITS) of the final fully connected layer. Let us refer to a given activation vector for a spectrogram input as v and an MNIST input as w.

Our goal is to map each v and w vector into a shared, h-dimensional space where matching image-audio pairs have high similarities. For our implemention, we set $h = 512$. As in [1], we start by computing $y = W_m v + b_m$ and $x = max(0, W_d w + b_d)$ with element-wise maximum. We then compute the score between image k and utterance l as $S_{kl} = max(0, y^T x)$. Thus, the S_{kl} is a non-negative single number. We optimize the alignment parameters $\theta = \{W_m, b_m, W_d, b_d\}$ using SGD

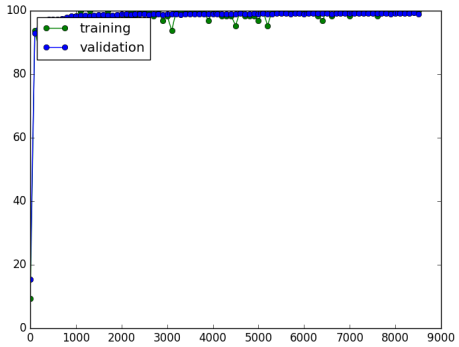


Figure 4: Percent validation accuracy of our MNIST convnet using 512-dimension fully connected layers over 10 epochs.

with a max margin objective function (hinge loss) that encourages higher scores for matching image-utterance pairs. We deviate from [1] by defining our cost as $C(\theta) = \sum_k \sum_l max(0, 1 - S_{kk} + S_{kl})$, where $k = l$ denotes the similarity score of a matching image-utterance pair.

We run stochastic gradient descent (SGD) to find an optimal θ configuration, defining a batch as a single randomly sampled MNIST image (retrieving its activation w from the MNIST convnet) with one utterance (retrieving its activation v from the spectrogram convnet) sampled from each of the 11 TIDIGITS classes. We use a learning rate of 1e-5 and run SGD for 500,000 iterations, initializing the θ parameters with a zero-mean Gaussian. The initial and final costs were 584,983.1 and 1,271.7, respectively.

4 Results

After training the MNIST convnet for 10 epochs, which totaled to be 8593 iterations, we achieved a 99.3% prediction accuracy on the single digit MNIST data. See Figure 4 for our results across iterations. For the spectrogram convnet, we achieved an accuracy of up to 89.3% using 512-dimension fully connected layers, and when using 1024-dimension fully connected layers we achieved up to 94.7% prediction accuracy on the validation set after 1500 iterations. We decided to use the 1024-dimension architecture in the alignment model due to its better performance. Figure 5 shows the performance of the two spectrogram convnet architectures over 1500 iterations.

To evaluate our alignment model, as in [1] we use our model to perform image annotation (in which we take an MNIST image and return the best matching utterance) and image retrieval (the opposite; we take a TIDIGITS utterance and return the best matching image) tasks. We use the tuned parameters $\theta = \{W_m, b_m, W_d, b_d\}$ from SGD in order to compute the matching score between image k (via its activation vector w) and spectrogram l (via its activation vector v). After deriving y and x as in Section 3.4, we modify the score S_{kl} slightly and define, for the evaluation phase alone, $S_{kl} = y^T x$, as we find that sometimes even the highest-scoring correct match has a negative score.

For the image annotation task, we randomly sample an MNIST digit k from our test set, and randomly sample a batch of one spectrogram from each of the 11 TIDIGITS classes. We compute

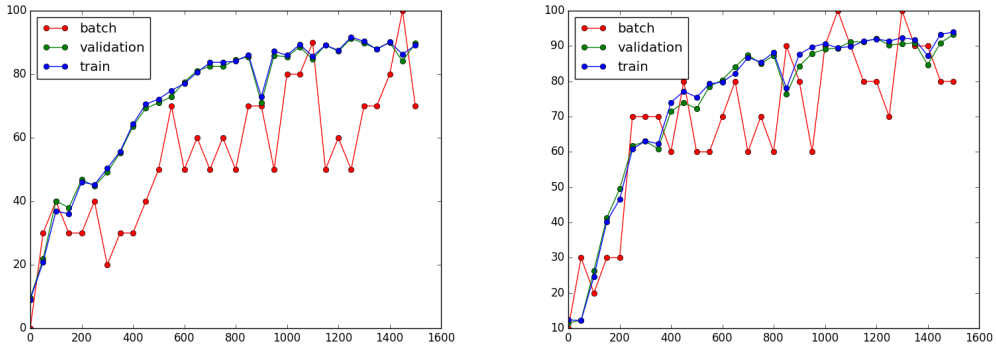


Figure 5: Percent validation accuracy of our spectrogram convnet using 512-dimension (left) and 1024-dimension (right) fully connected layers over 1500 iterations.

S_{kl} for each pair, and return the highest scoring l as our predicted match. We compare the ground truth labels to determine if our prediction was accurate. In order to determine percent accuracy, we repeated this process 100 times per MNIST class and report the image annotation accuracies in Figure 6. Overall our model achieved an image annotation accuracy of 88.5%.

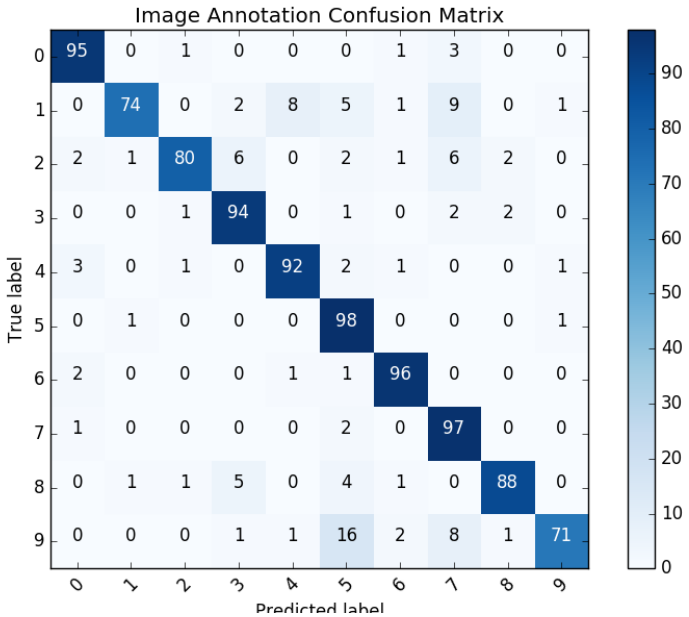


Figure 6: Confusion matrix of our Image Annotation Accuracy, comparing the top classification for each of the ten MNIST classes. 100 annotation tasks were performed per MNIST class, where the target image was randomly sampled from within the class and compared with a randomly sampled spectrogram from each of 11 TIDIGITS classes. We report an overall (average) accuracy of 88.5%. We combined the *oh* and *zero* labels to both represent the 0 digit in our performance evaluation.

The image retrieval task is similar. We randomly sample a TIDIGITS spectrogram l, and randomly sample a batch of one MNIST digit from each of the 10 classes. We compute S_{kl} for each pair, and return the highest scoring k as our predicted match. We compare the ground truth

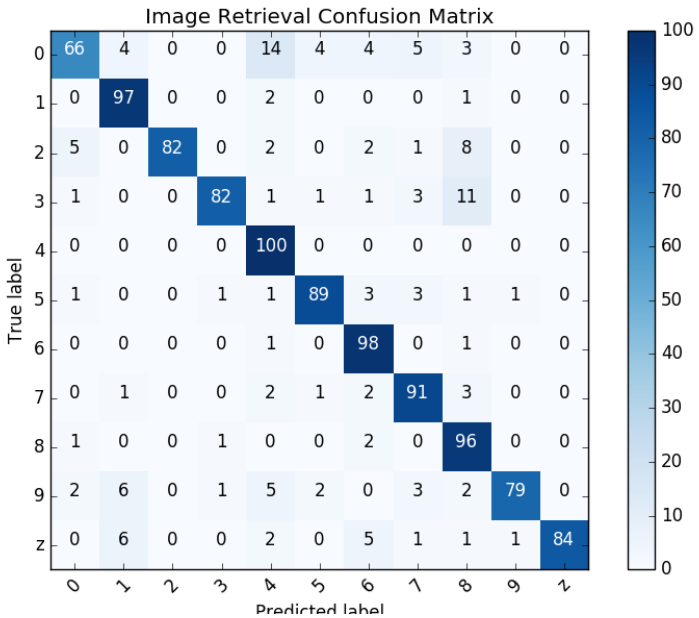


Figure 7: Confusion matrix of our Image Retrieval Accuracy, comparing the top classification for each of the eleven TIDIGITS classes. 100 annotation tasks were performed per TIDIGITS class, where the target spectrogram was randomly sampled from within the class and compared with a randomly sampled MNIST image from each of 10 MNIST classes. We report an overall (average) accuracy of 87.6%. Note that the 0 label represents the utterance *oh* whereas the z label represents *zero*.

labels to determine if our prediction was correct. As before, we repeated this process 100 times per TIDIGITS class and report the image retrieval accuracies in Figure 7. Overall our model achieved an image retrieval accuracy of 87.6%.

For both the annotation and retrieval tasks we explored, for the mismatches (if any), what were ground truth labels of the incorrectly predicted matches. The confusion matrices also show the classes of the incorrect predictions for annotation and retrieval tasks, respectively.

We can see that for the image annotation task, the MNIST digits 0, 5, 6, and 7 had the highest accuracies (at least 95%), whereas 1 and 9 had the lowest (under 75%). If we look at the common mismatches from Figure 6, we see that our model commonly outputted spectrograms for 5 and 7 incorrectly. On the other hand, the TIDIGITS spectrograms corresponding to labels 1,4, 6, and 8 had the highest accuracies of above 95%, while 9 and *oh* had the lowest (under 80%). With respect to mismatches, the MNIST digits 4 and 8 were commonly outputted incorrectly. While we can draw some conclusions in reasoning about the performance and mismatch behavior (like *oh* commonly mismatching with 4, 6, and 8 image representations because the digits have some visible similarity to 0), we need to more deeply investigate the behavior of the model with respect to particular image representations or spectrograms.

Overall, we believe that the accuracies shown in Table 6 and Table 7 demonstrate that our model is working successfully. Achieving overall accuracies of nearly 90% gives us confidence, as our model extremely outperforms a naive, random assignment model.

5 Conclusion

In this paper, we have described a model that can learn a joint semantic representation over MNIST images of digits and TIDIGITS spoken digit utterances. Our model, comprised of two convolutional neural networks tied together by an alignment model, aligns image representations of digits with their associated spoken representations, in the form of spectrograms. We run image annotation and image retrieval tasks in order to evaluate our model, reporting overall (average) accuracies of 88.5% and 87.6%, respectively. These results show promise for continued experimentation using neural models, and with TIDIGITS and MNIST data sets, for learning joint audio-visual semantic representations.

6 Future Work

There are several directions we plan to build on our research moving forward. We could better tune and experiment with variations of our convnets and our alignment model, trying to boost our annotation and retrieval accuracies. We may also experiment with variations of images or audio. For instance, we may include the TIDIGITS utterances from children instead of only adults to add variance to spectrogram forms, or insert noise into the MNIST images or distort the audio inputs to test the robustness of our initial model.

We plan to experiment with different training initializations of our convnets. For example, we might only train the MNIST convnet and randomly initialize the spectrogram convnet, or only train the spectogram convnet and randomly initialize the MNIST convnet. We also plan to experiment with an alignment approach similar to the one presented in [2], which, instead of having a separate alignment model to tie the networks together after they finish training on their respective data sets, inherently ties the networks together such that the θ parameters are optimized during training time. Namely, the networks are tied together by taking the dot products of the activations of their

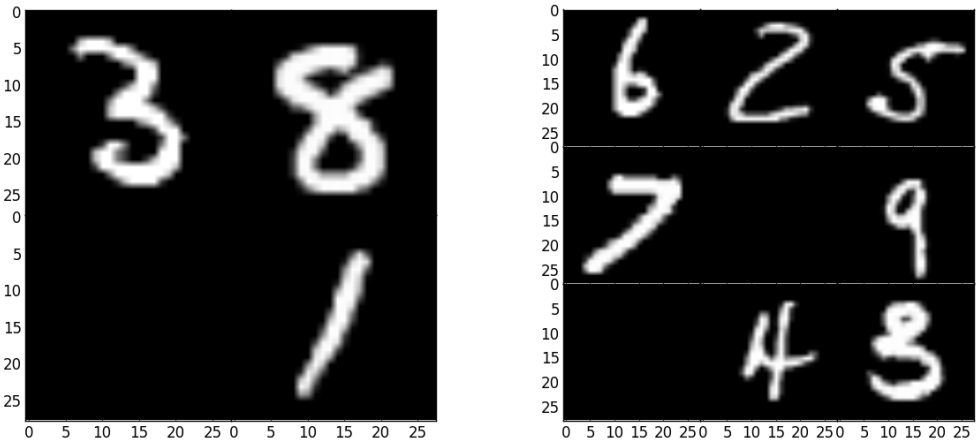


Figure 8: Randomized grid representations of MNIST digit sequences 138 (left) and 2345679 (right) with empty blocks inserted. This would be an interesting direction for future work.

penultimate layers (before the classification layer) and training the parallel convnet architecture using a custom cost function such that, at the end of training, θ is optimal.

One particular direction that we believe is a natural next step is dealing with sequences of digits rather than single digit utterances. Figure 8 presents two examples of sequence representations as randomized MNIST grids. We could alter our model and test its ability to associate spoken digits within particular regions of the image, perhaps using a region convolutional neural network (RCNN) for our image convnet as implemented in [1]. We could then analyze our results over a temporal dimension, investigating how our model matches digits in the MNIST grid to single digit utterances in a long (high number of frames) spectrogram of a digit sequence utterance. We had originally proposed to attempt this task as well but ran out of time due to some technical challenges in the single digit case.

Another question we plan to explore is: How will our model perform if the spoken digit utterances are in Spanish or another language rather than English? This opens up an interesting path towards machine translation, where we could attempt to train our model to learn not only joint semantic representations over images and spoken utterances in English but also make associations between the meanings of utterances across different languages.

Individual Contributions

Andrew Xia

Andrew helped Sitara with the re-rendering script to make the TIDIGITS audio files usable. He then worked on processing the spectrograms that Karan generated. He worked with Karan and Sitara to implement the spectrogram convnet, and took the lead on stripping the penultimate activation vectors from both convnets to be used in the alignment model. Andrew also worked on data organization, writing scripts to format data inputs to our model and ultimately analyzing the final results and creating the confusion matrix.

Karan Kashyap

Karan first worked on converting TIDIGITS audio files to spectrograms after they had been re-rendered. Following this, he experimented with various implementations of MNIST convnet architectures. He also helped Andrew and Sitara with the spectrogram convnet implementation. Lastly he trained the alignment model using Sitara's SGD script and wrote the evaluation code to obtain the final raw results. Karan also took the lead on writing the report, and wrote a script to generate grid representations of the MNIST digits, but the team ran out of time before being able to work with multi-digit sequences.

Sitara Persad

Sitara first worked on creating a script to re-render the TIDIGITS audio files into a usable format. She then worked with Andrew to process the spectrograms Karan generated. Sitara took the lead on the alignment model, implementing the SGD script and experimenting with momentum SGD and

various hyper-parameter settings to find an optimal configuration. She also worked with Andrew and Karan to implement the spectrogram convnet.

Acknowledgements

We would like to thank the 6.867 staff for their help throughout the semester and our TA Tuhin Sarkar, in particular, who helped us frame the project and overcome technical challenges in our project. We would also like to thank Jim Glass and David Harwath, whose work our project was based off, for encouraging us to take on this project and providing us with guidance.

References

[1] David Harwath, and James Glass. "Deep Multimodal Semantic Embeddings for Speech and Images." 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2015).

[2] David Harwath, Antonio Torralba, and James R. Glass. "Unsupervised Learning of Spoken Language with Visual Context" *30th Conference on Neural Information Processing Systems (NIPS 2016)*.

[3] Leonard, R. G., & Doddington, G. (1993). *TIDIGITS*. Retrieved December 1, 2016, from https://catalog.ldc.upenn.edu/ldc93s10.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.