GLU 2017 International Workshop on Grounding Language Understanding
25 August 2017, Stockholm, Sweden

Analysis of Audio-Visual Features for Unsupervised Speech Recognition

Jennifer Drexler, James Glass

MIT Computer Science and Artificial Intelligence Laboratory,
32 Vassar Street, Cambridge, MA, USA
jdrexler@mit.edu, glass@mit.edu

Abstract

Research on "zero resource" speech processing focuses on learning linguistic information from unannotated, or raw, speech data, in order to bypass the expensive annotations required by current speech recognition systems. While most recent zero-resource work has made use of only speech recordings, here, we investigate the use of visual information as a source of weak supervision, to see whether grounding speech in a visual context can provide additional benefit for language learning. Specifically, we use a dataset of paired images and audio captions to supervise learning of low-level speech features that can be used for further "unsupervised" processing of any speech data. We analyze these features and evaluate their performance on the Zero Resource Challenge 2015 evaluation metrics, as well as standard keyword spotting and speech recognition tasks. We show that features generated with a joint audio-visual model contain more discriminative linguistic information and are less speaker-dependent than traditional speech features. Our results show that visual grounding can improve speech representations for a variety of zero-resource tasks.

1. Introduction

Unsupervised speech processing is a problem that has seen increasing interest in recent years [1, 2], owing largely to the expense of collecting the text annotations required to build supervised speech recognition systems. While most recent unsupervised speech processing work has used only speech data [1, 3, 4], there has been some early work by Roy [5] that considered both audio and visual modalities for language learning. This "sensor-based" language learning [2] framework uses visual grounding to help constrain the language learning problem.

Recently, Harwath et al. have proposed a model that is able to learn a latent audio-visual embedding space from pairs of images and corresponding spoken descriptions [6]. Using this weak form of contextual grounding, they demonstrate that the embedding vector is able to learn word-like units [7]. Their results inspire us to examine whether the internal representations learned by the speech processing component of their model are effective for standard unsupervised speech processing tasks. In this work, we take only the trained audio-processing branch of this joint audio-visual model and use it to generate features from other audio corpora (with no associated images) that have been used to evaluate previous "zero-resource" research. This allows us to make comparisons with speech-only representations on these tasks, to assess the potential benefit of the audio-visual latent representation for language learning.

As a preface to these downstream experiments, we perform both qualitative and quantitative analyses of the features generated at each level of the network. Through these analyses, we seek to understand the inner-workings of this audio-visual network - specifically, the nature of the receptive fields of the units at each layer of the network and what kind of information (e.g. phonetic, semantic, speaker) those units represent.

2. Prior Work

This paper directly follows from [6] and [7]. In those papers, Harwath et al. trained two convolutional neural networks (one for audio and one for images) to project images and their spoken captions into the same embedding space. Using these embeddings, the authors were able to perform effective image search and annotation [6], as well as discover clusters of word-like units within the audio captions that correlate with specific visual concepts [7].

Our analysis of the audio-visual network is inspired by prior work in computer vision; that field is largely reliant on convolutional neural networks and has thus developed a variety of techniques for analyzing their inner workings. We adopt the analyses used by Zhou et al. [8], particularly their visualizations of the receptive fields of individual units within a network.

We judge the usefulness of audio-visual features for downstream speech applications using pre-existing application and evaluation code from three sources. First, we compare the audio-visual features to submissions from the first track of the 2015 Zero Resource Challenge [9]. We compare against two high-performing submissions from the Challenge - [10] and [11]. Both models start by performing unsupervised term detection using dynamic time-warping. This produces frame-level alignments, which can then be split into pairs of aligned frames that should contain the same linguistic information. Both [10] and [11] develop neural network architectures designed to learn a representation space which brings aligned frames together and pushes unaligned frames apart.

The second level of downstream analysis performed here is spoken term detection. We use the TIMIT keyword spotting task introduced in [3]. In that work, Lee and Glass describe an unsupervised Bayesian nonparametric model which segments and clusters speech into phone-like units. The features they use for keyword spotting are frame-level posteriorgrams over those discovered units. That model continues to be the best performing fully unsupervised model applied to this task. However, Harwath et al. [6] showed significantly improved performance on this task with the addition of visual information.

Finally, we use the unsupervised speech recognition system developed in [4] to evaluate the audio-visual features' usefulness in a full speech recognition task. Kamper et al. [4] develop a model that is similar to [3] but performs segmentation and clustering of speech into word-like, rather than phone-like, units. The authors judge their model's performance with unsupervised phone and word error rate metrics. Kamper et al. experiment with both MFCC features and correspondence autoencoder (CAE) features [11]; their results suggest the potential of

10.21437/GLU.2017-12

this type of model but show the need for additional research to produce usable unsupervised speech transcripts.

3. Data and Methods

3.1. Audio-Visual Model Architecture

In this section, we briefly describe the audio-visual model used to generate features for experiments. The model is described in more detail in [7].

The audio-visual model has two branches: one that processes audio input, and one that processes images. The image branch, an off-the-shelf VGG network [12] trained on labeled images, will not be discussed here. It is sufficient to note that the output layer of the network is replaced by a fully connected layer that projects the last activation layer into a 1024-dimensional embedding space.

The input to the audio network is 40-dimensional log Mel filterbank features computed with a 25ms window and 10ms frame shift. The first layer of the network is a convolution over a single frame. Each subsequent convolutional layer (for a total of five) has an increasingly larger filter width. Each of these convolutions is followed by a ReLU nonlinearity, and a max-pool operation that halves the frame rate. The output of the final layer is meanpooled over the entire utterance to form a single 1024-dimensional embedding of each audio caption.

We compute the similarity between an image and a spoken caption by taking the dot product of their embeddings. The network is trained to maximize the similarity between matching image/caption pairs while minimizing the similarity between mismatched pairs. These mismatched pairs are generated by randomly sampling one impostor image/caption pair for each true image/caption pair in the training set.

3.2. Data

The audio-visual model was trained using a dataset of images taken from the Places205 dataset [13], each paired with several spoken captions collected via Amazon Mechanical Turk. See [6] for additional data collection details. We use the Google Speech API[1] to generate transcripts of the audio captions; we do not have access to gold transcripts or alignments. We will refer to this corpus as the Places dataset for the remainder of this paper.

We evaluate our audio-visual features on two standard speech corpora with no associated images: TIMIT [14] and the Buckeye Corpus [15].

3.3. Feature Generation

For the experiments in this paper, we took the trained audio branch of the audio-visual model, fixed the weights, and used it to generate features for new audio which it had not seen during training. We extract features after each convolution and nonlinearity in the model, resulting in five feature sets, labeled in order from the lowest level of the hierarchy (AVNet1) to the highest (AVNet5). These feature representations range in dimensionality from 128 features to 1024 and in receptive field size from 25ms to almost 2 seconds.

For the feature analysis, keyword spotting experiments, and Zero Resource Challenge evaluation, we directly used the features extracted from the network. For unsupervised speech recognition, all frames were first PCA-whitened.

[1] https://cloud.google.com/speech/

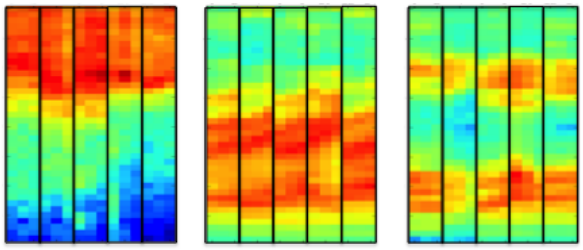

Figure 1: *Filterbank features for TIMIT segments that most activate three different neurons in the first AVNet layer. (a) sh, jh, sh, z, z (b) r, er, r, r, r (c) l, l, w, l-w, ow.*

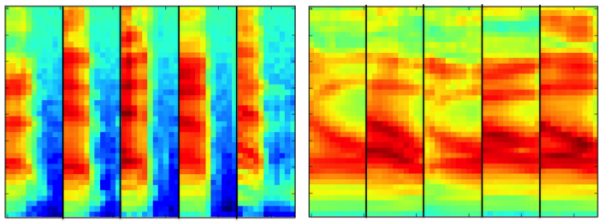

Figure 2: *Filterbank features for TIMIT segments that most activate two different neurons in the second AVNet layer. (a) ae-tcl (that), eh-tcl (that), ae-tcl (that), ae-pcl (aptitude), aa-tcl (cannot) (b) eh-l (hotels), aw (how), ao-l (all), aa-l (doll), ae-l (alfalfa).*

4. Feature Analysis

Following [8], we visualize the activation pattern of an individual unit in our network by looking at the speech segments that most activate that unit. We use TIMIT for this analysis, so that we can associate each segment with its corresponding phoneme label(s). Figures 1 and 2 show the TIMIT speech segments that most activate three units from AVNet1 and two units from AVNet2, respectively.

In both cases, the similarities between the speech segments associated with a given unit are clear from visual inspection. The AVNet1 units seem to select for a particular phoneme - of the top 100 segments that activate each unit at this level, on average 37.8 are instances of the same phoneme, and 61.4 are from the same broad phone class. AVNet2 units have a larger receptive field, and seem to select for a particular bi-phone. Of the top 100 segments that activate each unit at this level, on average 17.2 are instances of the same phoneme sequence, and 37.4 are from the same sequence of phone classes.

Figure 3 shows the three segments from the TIMIT data that most activate one unit in the third layer. While these segments look quite different on immediate visual inspection, they share many similarities. Each one starts with a vowel and strong fricative, and they all end with the phoneme sequence “pcl-p-r-iy”. At this layer, we can see that the network has learned to be invariant to the way different speakers pronounce similar sequences of sounds.

Table 1 shows results of a quantitative analysis of the receptive fields of the units in the network, using the Places data. At each level, we form a ”cluster” for each unit by selecting the 100 speech segments that most activate that unit. We assign each cluster a speaker and word label based on the most common speaker or word in those 100 segments (ignoring common

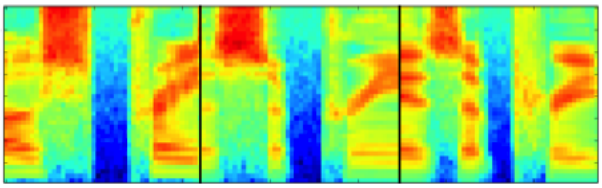

Figure 3: *Filterbank features for TIMIT segments that most activate one neuron in the third AVNet layer. (a) 'ow', 's', 'pcl', 'p', 'r', 'iy' ([m]ost pre[cincts]), (b) 'ax', 's', 'ax', 'pcl', 'p', 'r', 'iy' (a supre[me]), (c) 'ih', 'z', 'ix', 'pcl', 'p', 'r', 'iy' (is appre[ciated]).*

Table 1: *Average word and speaker purity of the clusters composed of the top 100 speech segments from the Places data that activate each unit at a given layer of the network.*

Features	Speaker Purity	Word Purity
AVNet1	0.269	0.045
AVNet2	0.212	0.052
AVNet3	0.134	0.129
AVNet4	0.112	0.198
AVNet5	0.101	0.207

stop-words). These labels can then be used to calculate speaker and word purity. As Table 1 shows, units in the higher layers of the network are less-speaker dependent (lower speaker purity) and more representative of word-level information.

Based on these results, we can hypothesize that the lower layers of the network should be useful for low-level tasks like the Zero Resource Challenge evaluation, while higher-level features may be more useful for keyword spotting and unsupervised speech recognition.

5. Zero Resource Challenge

5.1. Evaluation

The first track of the Zero Resource Challenge is evaluated using a minimal pair ABX task inspired by a standard paradigm used in psychophysics experiments. This task judges a feature representation by computing, over a large dataset, the likelihood that any given phoneme segment (X) is closer to another instance of the same phoneme (A) than an instance of a different phoneme (B). See [9] for full evaluation details.

The evaluation is divided into two conditions: within speaker (where A, B, and X all come from the same speaker) and across speaker (where A and B come from one speaker and X from another). The default distance metric provided is cosine distance, which we also use here.

5.2. Results

Table 2 shows results of the Zero Resource Challenge ABX evaluation. Of the audio-visual features, the best results came from the AVNet2 features. Both the AVNet1 and AVNet2 features are better than filterbank features on both conditions, while AVNet3 features are better than the filterbank features on the across-speaker condition. This supports our hypothesis that the use of visual information can reduce speaker effects and make cross-speaker learning easier.

Despite these encouraging results, prior work using CAE [11] and ABNet [10] models to generate features generated better results than our AV features. However, these models both rely on term detection systems that find repeated patterns in speech from simple MFCC features. While such experiments are beyond the scope of this paper, the results presented in the following section suggest that using audio-visual features as input to those algorithms could improve term detection across speakers, which could in turn improve the performance of the CAE and ABNet features.

Table 2: *Results of 2015 Zero Resource Challenge ABX English evaluation, by input feature type. Cosine distance was used with all features.*

Features	Within	Across
Filterbank	16.3	29.7
CAE [11]	13.5	21.1
ABNet [10]	12.0	17.9
AVNet1	15.8	26.9
AVNet2	**14.4**	**23.5**
AVNet3	20.2	27.8
AVNet4	27.6	34.8
AVNet5	35.1	40.0

The features from AVNet4 and AVNet5 are significantly worse than filterbank features on both the within-speaker and across-speaker ABX tasks. This is not surprising: if these layers are, in fact, representing higher-level linguistic information (and perhaps even semantic information), we would expect them to perform poorly on a phone-level discrimination task.

6. Keyword Spotting

For our first spoken term detection task, we use the AVNet features on the TIMIT keyword spotting task defined in [3]. Harwath and Glass [6] reported results on this task using the first layer of a similar AV network with a slightly different architecture. As in that work, we find that the AVNet features (in this case, from AVNet2) are able to match prior work in terms of search precision while significantly outperforming that prior work in terms of equal error rate. Perhaps surprisingly, given the prior demonstration of the ability to discover word clusters using the high-level AVNet features [7], the higher layers of the network are not at all useful for this task. However, Harwath et al. [7] showed an ability to discover words associated with visual concepts - in contrast, the words used for this keyword spotting task are not visual.

We develop two keyword spotting tasks from the Places dataset based on the hypothesis that the last layers of the network will do a good job of representing words associated with specific visual concepts, but will not effectively differentiate between non-visual words. For this task, we chose words that occur more than 15 times in the Places development set and more than 10 times in the test set. For each keyword, we select the first 15 instances from the development set as queries, so the system will have equal exposure to all keywords in both sets. Scores are calculated over all instances in the test set. Table 3 lists the words selected for these two tasks (as well as the keywords from the TIMIT task for comparison).

Table 4 shows results in terms of precision at N (P@N) of using the different AVNet features for keyword spotting tasks. In accordance with our hypothesis, we find that the higher lay-

Table 3: *Keywords for keyword spotting tasks.*

Task	Keywords
TIMIT	development, organizations, money, age, artists, surface, warm, year, problem, children
Places (non-visual)	about, across, appears, background, bottom, different, everything, filled, foreground, maybe, mostly, nearby, picture, possibly, probably, several, something, underneath, very, visible
Places (visual)	building, cabinet, camera, classroom, construction, flower, fountain, garden, hospital, mountain, outside, patio, people, railroad, river, subway, table, water, windmill, window, woman, wooden

Table 4: *Keyword spotting results. Average precision at N (P@N) across all task keywords.*

Task	MFCC	AV1	AV2	AV3	AV4	AV5
TIMIT	50.0	52.7	**62.0**	58.1	39.3	15.8
Non-visual	16.8	24.3	47.1	**51.2**	45.0	23.2
Visual	13.4	19.3	43.9	56.5	**63.5**	**63.2**

ers of the network are very effective for representing words associated with visual information but ineffective for non-visual words. These results highlight the importance of the training objective in determining how information is represented in the network. They suggest that we might get improved results on downstream tasks by adding a secondary objective to this network - we want to use visual information to help learn relevant invariances in speech, but we also want our the embeddings to effectively represent all words in an utterance.

7. Unsupervised Speech Recognition

7.1. Model

For this evaluation, we borrow the unsupervised speech recognition model developed in Kamper et al. [4]. Our main contribution is to test the effectiveness of using audio-visual features as input. We follow their lead on all of the parameter settings except those outlined in the following paragraphs.

For MFCC features, as in [4], we used 13 dimensions per frame and create fixed-length segment embeddings by downsampling the data to 10 frames per segment. Because of both the high dimensionality of the audio-visual features and the fact that higher AVNet layers cover larger windows of speech, we pursued more aggressive downsampling at successively higher layers: AVNet1 was downsampled to 11 frames per segment, AVNet2 to nine, AVNet3 to seven, AVNet4 to five, and AVNet5 to three. For all audio-visual features, we reduced the covariance scale of the discovered clusters to $1e-5$ (we used $1e-3$ for the MFCC features, as in [4]).

Finally, we replaced the initial iterations of acoustic model sampling with a simpler k-means cluster assignment. We found this to be more efficient for our higher-dimensional representations without reducing model performance.

7.2. Evaluation

As in Kamper et al. [4], we evaluate our unsupervised speech recognition system on the Buckeye corpus. Kamper et al. [4] focus on word error rate (WER); because we are not explicitly hoping to discover word-level units, we also evaluate phone error rate (PER).

7.3. Results

Table 5: *Unsupervised recognition results on the Buckeye corpus, using the system described in [16].*

Features	WER	PER	Num Segments
MFCC	86.2	71.2	24871
MFCC mindur	83.6	76.1	15522
AVNet1	85.0	73.0	22735
AVNet2	82.9	**69.3**	23135
AVNet3	**82.0**	69.9	20865
AVNet4	86.0	78.2	19836
AVNet5	91.7	93.0	18123

Table 5 shows unsupervised many-to-one WER and PER using MFCC features or audio-visual features. The best overall WER comes from the AVNet3 features, while the best PER comes from the AVNet2 features. Both AVNet2 and AVNet3 features improve on the MFCC results on both measures. While Kamper et al. [4] were able to improve their results by imposing a heuristic minimum segment duration, we are able to achieve better results without such a constraint.

The most obvious effect of the minimum duration constraint is to force the model to discover a reduced number of segments. The last column of Table 5 shows that we can produce a similar effect using subsequent layers of the audio-visual model.

8. Conclusions

In this paper, we showed how information about speech is represented in a network trained to project audio captions into an embedded feature space near their associated images. We saw that units in the lowest layers are tuned to specific phonetic classes, those in the middle layers are selective for both phonetic and semantic information, and those at the highest layers are almost exclusively semantic. Additionally, we showed a increase in speaker-invariability as we move up through the network.

Along with these unit-level analyses, we explored the use of these features for a variety of unsupervised speech processing tasks. We showed specifically that the semantic nature of the representations in the highest layers of the network render those features ill-suited for unsupervised speech processing tasks, unless those tasks take specific advantage of those semantics (e.g. spoken term detection for terms associated with visual concepts). Nonetheless, the high-level semantic task for which this network was trained is an effective means of learning phonetic representations and speaker invariance in the lower levels of the network. We showed that these low-level features are able to outperform baseline filterbank or MFCC features in out-of-domain spoken term detection and unsupervised speech recognition tasks. Overall, these results suggest that grounding speech in a visual context can significantly improve the performance of zero-resource speech processing.

9. References

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

[2] J. Glass, "Towards unsupervised speech processing," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*. IEEE, 2012, pp. 1–4.

[3] C.-y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[4] H. Kamper, A. Jansen, and S. Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *arXiv preprint arXiv:1606.06950*, 2016.

[5] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive science*, vol. 26, no. 1, pp. 113–146, 2002.

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

[7] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," *to appear ACL 2017; arXiv preprint arXiv:1701.07481*, 2017.

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.

[9] M. Versteegh, R. Thiolliere, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015." in *INTERSPEECH*, 2015, pp. 3169–3173.

[10] R. Thiolliere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling." in *INTERSPEECH*, 2015, pp. 3179–3183.

[11] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge." in *INTERSPEECH*, 2015, pp. 3199–3203.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[14] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The darpa timit acoustic-phonetic continuous speech corpus cdrom," *Linguistic Data Consortium*, 1993.

[15] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," *Columbus, OH: Department of Psychology, Ohio State University*, 2007.

[16] H. Kamper, A. Jansen, and S. Goldwater, "Fully unsupervised small-vocabulary speech recognition using a segmental bayesian model." 2015.