INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Transfer Learning from Audio-Visual Grounding to Speech Recognition

Wei-Ning Hsu, David Harwath, James Glass

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA
{wnhsu,dharwath,glass}@mit.edu

Abstract

Transfer learning aims to reduce the amount of data required to excel at a new task by re-using the knowledge acquired from learning other related tasks. This paper proposes a novel transfer learning scenario, which distills robust phonetic features from grounding models that are trained to tell whether a pair of image and speech are semantically correlated, without using any textual transcripts. As semantics of speech are largely determined by its lexical content, grounding models learn to preserve phonetic information while disregarding uncorrelated factors, such as speaker and channel. To study the properties of features distilled from different layers, we use them as input separately to train multiple speech recognition models. Empirical results demonstrate that layers closer to input retain more phonetic information, while following layers exhibit greater invariance to domain shift. Moreover, while most previous studies include training data for speech recognition for feature extractor training, our grounding models are not trained on any of those data, indicating more universal applicability to new domains.
Index Terms: transfer learning, audio-visual grounding, multimodal learning, semantic supervision, speech recognition

1. Introduction

Robustness of automatic speech recognition (ASR) systems is essential to generalization of using speech as interfaces for human computer interaction. Thanks to the strong modeling capacity of neural networks, recent studies [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] have demonstrated that by providing supervised examples as abundant and diverse as possible, such models can learn to extract domain invariant features and recognize linguistic units jointly. However, without additional treatment, good performance and robustness may not be achieved when labeled data are very limited in quantities or not available in all domains [11]. One way to ease the burden of ASR systems is by providing better features, which are more invariant to nuisance factors while containing linguistic information ready for use (e.g., linear separability w.r.t. phonemes). Such features can be hand-crafted by leveraging prior knowledge [11, 12, 13, 14], or they can be learned in a data-driven fashion. Furthermore, this learning can take place jointly with ASR [15], or separately with some tasks that have aligned objectives [16, 17, 18].

Learning features from some source tasks that can benefit the target task is a common realization of transfer learning [19]. In this work, we propose a novel inductive transfer learning scenario [19], which utilizes speech features learned from audio-visual grounding for speech recognition. Audio-visual grounding [20] is a task which aims to distinguish whether a spoken caption is semantically associated with an image or not, and vice versa, without using any textual transcripts. Deep audio-visual embedding network (DAVEnet) [21] is a two-branched convolutional neural network model for this task, which learns to encode images and spoken captions into a shared embedding space that reflects semantic similarity. To successfully learn a semantic representation for speech, the model has to recognize its lexical content, which in turns requires identifying phonetic content. Therefore, one can expect intermediate layers of the speech branch in DAVEnet models to function as lexical or phonetic unit detectors. Furthermore, since non-linguistic aspects of speech, such as speaker, are not correlated with semantics, these information may be discarded, resulting in the intermediate outputs from the model being invariant to domain shift.

We conduct a series of ASR experiments probing properties of the features distilled from DAVEnet models at different layers. Results indicate higher in-domain accuracy using features closer to input, and better robustness to domain shift using features from latter layers. In addition, we also study how the choice of DAVEnet architectures and grounding performance affects the performance of distilled feature extractors. In summary, our contributions are three-fold: (1) To the best of our knowledge, this is the first work connecting audio-visual grounding with speech recognition. (2) Our empirical study verifies that the distilled feature extractors not only contain sufficient information for recognizing phonemes, but better remove nuisance information. (3) Moreover, the grounding models are trained on a different dataset from that used for ASR, indicating more general applicability of the distilled features.

2. Learning Spoken Languages through Audio-Visual Grounding

In this section, we describe in detail the source task as well as the DAVEnet model, and then review several analysis studies which lay the foundation for our work.

2.1. Audio-Visual Grounding

Inspired by the fact that humans learn to speak before being able to read or write, audio-visual grounding of speech is a proxy task proposed in [20] that aims to examine the capability of computational models to learn a language using only semantic-level supervision. To simulate such a learning scenario, a model has access to images and their spoken captions during training. The goal of the model is to learn a semantic representation for each caption and each image, such that representations of semantically correlated utterances and images are similar to each other, while those from irrelevant pairs are dissimilar. Performance is evaluated using a cross-modality retrieval task: given a spoken sentence, a model is asked to rank a list of 1,000 images according to semantic relevance, with only one image being the correct answer, and vice versa. Recall@10 averaged over the retrieval tasks in both directions is used for evaluation.

http://dx.doi.org/10.21437/Interspeech.2019-1227

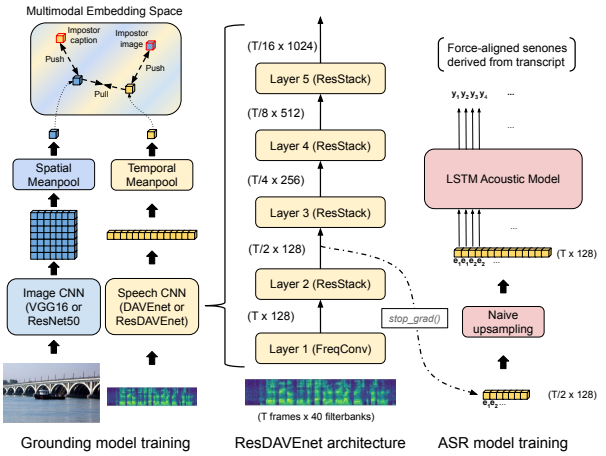

Figure 1: *Graphical illustration of audio-visual grounding model training (left), ResDAVEnet architecture (center), and feature distillation pipeline for speech recognition (right).*

2.2. Deep Audio-Visual Embedding Network (DAVEnet)

DAVEnet is a convolutional neural network (CNN) for audio-visual grounding proposed in [20, 22, 21], which consists of two branches: f for speech and g for image, as depicted in Figure 1. Each branch has a sequence of strided convolutional blocks, followed by a global mean-pooling layer to produce a fixed dimensional representation. The model is trained to minimize a triplet loss [23, 24]: given a similarity measure $S(\cdot, \cdot)$, paired speech and image, x_s and x_i, along with one imposter instance from each modality, $\tilde{x}_s$ and $\tilde{x}_i$, the loss enforces $S(f(x_s), g(x_i))$ to exceed both $S(f(x_s), g(\tilde{x}_i))$ and $S(f(\tilde{x}_s), g(x_i))$ by a predefined margin. Following [25], imposter instances are drawn using a mixture of uniform sampling and within-batch semi-hard negative mining [24]. $S(z_1, z_2) = \langle z_1, z_2 \rangle$ is used here.

In our experiments, we make use of two DAVEnet model variants. The first is identical to the model used in [21], which uses an audio network comprised of 5 convolutional layers and the VGG16 architecture for the image network. The second model, ResDAVEnet, is based upon deep residual networks [26]. The image network makes use of the ResNet50 architecture, while the audio network is based on strided 1-D convolutions with residual connections. The first layer of the ResDAVEnet audio model is comprised of 128 convolutional units each spanning all frequency channels but only one temporal frame, with a temporal shift of 1 frame. This is followed by a ReLU and a BatchNorm layer. The remainder of the network is a sequence of 4 residual stacks with channel dimensions 128, 256, 512, and 1024. Each residual stack is comprised of a sequence of two basic residual blocks (as described in [26]) which share the same overall channel dimension, with 2-D 3x3 kernels replaced with 1-D kernels of length 9. Additionally, the first residual block in each layer in each stack is applied with a stride of two frames, resulting in an effective temporal downsampling ratio of 2^4 for the entire network, as shown in Figure 1 (center).

2.3. Emergence of Multi-Level Speech Unit Detectors

Recent work [27, 22, 25] on analyzing DAVEnets have shown that, despite the fact that phoneme and word labels are never explicitly provided, such detectors automatically emerge within these models. Phoneme-like detectors reside in layers closer to the input [27, 25], while semantic word detectors reside in layers closer to the output [22]. Such findings echo with the recent discovery in the computer vision community [28, 29] that in a trained scene classifier, layers closer to sensory input appear to be low-level pattern (e.g., shape, edge, and color) detectors, while object detectors emerge at later layers. This behavior can be mainly attributed to the compositionality of the prediction target as well as the inductive bias we impose in the model architecture. Just as a scene can often be determined by the objects that are present, the semantics of a spoken sentence is determined by the sequence of words, each of which in turn is determined by phoneme sequences. Prediction of semantic objects from a spoken sentence can therefore be regarded as a bottom-up process, which iteratively composes higher-level concepts from lower-level ones with the hierarchical convolution operations in CNNs.

3. Transfer Learning to Speech Recognition

3.1. Distilling Robust Feature Extractors for ASR

Both DAVEnet variants are trained on the Places Audio Caption dataset (PlacesAudCap) [21], derived from the Places205 scene classification dataset [28]. PlacesAudCap is composed of over 400K image and unscripted spoken caption pairs collected from 2,954 speakers via Amazon Mechanical Turk, which sums up to over 1,000 hours. For the audio-visual grounding task, both models use 40-dimensional log Mel filterbank (FBank) features with 10ms shift and 25ms analysis window as input, and achieve R@10 of 0.629 and 0.720, respectively.

As a natural result of large-scale crowd-sourcing, this dataset exhibits great diversity not only in textual content, but also in speaker, background noise, and microphone channels. For both semantic grounding and speech recognition, these non-textual factors are nuisances to the target, and therefore would eventually be removed from the internal representations learned by the networks trained for these two tasks. Having been exposed to a vast amount of nuisance factors, we hypothesize that the audio branch of DAVEnet models would also learn domain invariant phonetic representations at later layers, which can be subsequently utilized for robust speech recognition. From now on, we denote features extracted from the k-th layer of model M with M-Lk, for example, ResDAVEnet-L2.

To account for the different frame rates at different layers in DAVEnet models, when extracting outputs from a layer with a down-sampling rate r compared to the speech inputs, we repeat each step r times for simplicity, as shown in Figure 1 (right).

3.2. Evaluating Transfer Learning Performance

To evaluate transfer learning performance, we consider three criteria: (1) inclusion of phonetic content, (2) exclusion of nuisance factors, and (3) transferrability across datasets. The first two are evaluated using a protocol similar to [17], where an ASR model is trained on a set of domains, and evaluated on both in-domain and out-of-domain speech (relative to the training data). Performance on in-domain data characterizes an upper bound for the amount of phonetic information that can be inferred from the input. The performance gap between in-domain and out-of-domain data quantifies the invariance of the features to nuisance factors: the smaller this gap, the more invariant the features are. To test the third criteria, instead of training the source task on a dataset that includes speech used for the target task, a separate dataset collected through a different process (i.e., PlacesAudCap) is used. We emphasize here that this is a

more practical setting to consider than training one feature extractor for each target task.

4. Related Work

Transfer learning has a long history in the field of machine learning [19]. More recently, deep neural network models have been shown to be extremely effective for learning representations of data with a high degree of re-usability across many different tasks and domains. Perhaps the most well-known example of this is the use of the ImageNet [30] image classification database to pre-train convolutional neural network models for other downstream computer vision tasks [31, 32, 33]. Other sub-fields have also developed similarly techniques. For example, in natural language processing, dense word vector models such as word2vec [34] and GloVe [35], or more advanced ones like ELMo [36] and BERT [37] have quickly replaced one-hot word representations in many tasks and pushed the state-of-the-art forward on a variety of language understanding tasks. More recently, there is also an increasing interest in learning from multimodal data [38] and transfer learned representations from such tasks [39]

In the field of speech recognition, low-resource speech recognition is a scenario which heavily benefits from transfer learning, for example in the form of training on multilingual datasets [40]. Other models capable of disentangling phonetic and domain information have recently been shown to learn acoustic features with a greater degree of domain invariance than traditional acoustic features [16, 7, 17]. Another line of work has studied the use of the visual modality as a form of weak supervision using semantic information for acoustic modeling [20, 41, 42], followed up with analysis on representations learned from such models [27, 43, 25]. In this paper, we build upon this prior work and quantify the degree to which these representations can be used to build robust ASR.

5. Experiments

5.1. ASR Setup and Baselines

We consider TIMIT [44] and Aurora-4 [45] for training ASR systems to study robustness of the proposed method to speaker, channel, and noise. TIMIT contains 5.4 hours of 16kHz broadband recordings of read speech from 630 speakers, of which about 70% are male. Recordings from male speakers are used for training ASR systems, which are then tested on both genders. Aurora-4 is based on the Wall Street Journal (WSJ) corpus [46], containing recordings with microphone and noise variation. The set of conditions are divided into four groups: clean (A), noisy (B), channel (C), and noisy+channel (D). While recordings in A are recorded by one microphone in quiet environments, those in C are recorded with a different set of microphones than A. Recordings in B and D are created from A and C, respectively, with artificially added noises. Similar to [17], we use the clean set (A) for training ASR systems, and test on the four groups separately.

Kaldi [47] is used for training of initial HMM-GMM models, forced alignment, and decoding. The Microsoft Cognitive Toolkit (CNTK) [48] is used for neural network-based acoustic model training. To simplify the pipeline and study only the effect of ASR input features, the same forced alignment derived from a HMM-GMM model trained on Mel-frequency cepstral coefficient (MFCC) features are used for all experiments, following the default recipe in Kaldi. A three-layer long short-term memory (LSTM) acoustic model with 1,024 memory cells and a 512-node linear projection at each layer is used [49]. Training of LSTM acoustic models closely follows [50], which minimizes a frame-level cross-entropy loss using stochastic gradient descent with a momentum of 0.9 starting from the second epoch. Initial learning rate is set to 0.2 per minibatch, and $L2$ regularization with a weight of $1e-6$ is used.

We consider two types of features to compare with our proposed method. The first one is FBank feature, which is the input to DAVEnet models and contains rich phonetic and domain information. The second one is the latent segment variable z_1 from a model called factorized hierarchical variational autoencoder (FHVAE) [16]. FHVAE learns to encode sequence-level and segment-level information into separate latent variables without supervision by optimizing an evidence lower bound derived from a factorized graphical model, and has been shown effective for extracting domain invariant ASR features [17].

While previous work investigated usage of FHVAE for ASR by training FHVAE models on all domains of the target task (e.g., Aurora-4 with all four conditions) [17, 8], we also evaluate FHVAE models trained on PlacesAudCap to test cross-dataset transferability, and on the subset of domains used for ASR training. We use FHVAE models with two LSTM layers, each with 256 cells, for both the encoders and decoder. A discriminative weight of $\alpha = 10$ is applied for all models, and the scalable training algorithm proposed in [51] is used for training on PlacesAudCap dataset with a sequence batch size $K = 5000$, because the original algorithm cannot handle large-scale datasets.

5.2. Main Results

Tables 1 and 2 present the testing word error rates (WERs) on both in-domain and out-of-domain conditions for ASR systems trained with different features. *FE Train Set* denotes the data used for training feature extractors, and *A/I* following *Places* represents the audio and image portion of the PlacesAudCap dataset, respectively.

Starting with Table 1, we observe that FBank suffers from severe degradation in all out-of-domain conditions (B, C, and D), while FHVAE trained on all conditions of the Aurora-4 dataset achieves the best performance. However, when trained on *Places A*, improvement of FHVAE from FBank on out-of-domain data becomes less significant in the presence of additive noise, compared to the result in the purely channel-mismatched condition (C). Results of the proposed methods are shown in the second and the third section in Table 1. While features from ResDAVEnet consistently outperforms FBank and FHVAE (Places A) for all layers, those from DAVEnet do not. We hypothesize that the much deeper architecture of ResDAVEnet at each layer (ResStack) enables better removal of nuisance factors and preserving of linguistic information compared to DAVEnet, which also reflects in the comparison of grounding performance as mentioned earlier.

It is also worth noting that, for both DAVEnet and ResDAVEnet models, performance in matched domain degrades when using latter layers, and except for ResDAVEnet-L1, all features are actually worse than the FBank baseline. This could indicate discarding of relevant phonetic information in the process of inferring higher-level semantic representation such as words. Table 2 demonstrates a similar trend as Table 1, where FHVAE trained on TIMIT dataset of all genders achieves the best out-of-domain WER, and ResDAVEnet-L2 is the best comparing to models trained on Places.

We also present qualitative visualizations in Figure 2 using

Table 1: *Aurora-4 test WERs for different ASR features. A is the domain matched with the ASR training set.*

ASR Feature	FE Train Set	Test WER (%) Avg.	A	B	C	D
FBank	N/A	53.38	4.02	50.77	40.13	66.31
FHVAE	Places A	49.31	4.37	44.43	26.64	65.33
DAVEnet-L1	Places A+I	57.89	3.40	54.92	46.89	71.69
DAVEnet-L2	Places A+I	57.05	4.56	56.15	34.88	70.35
DAVEnet-L3	Places A+I	61.65	8.52	60.53	35.57	75.90
ResDAVEnet-L1	Places A+I	44.03	**2.91**	38.53	36.86	57.53
ResDAVEnet-L2	Places A+I	**33.11**	4.20	**25.17**	27.09	46.75
ResDAVEnet-L3	Places A+I	33.16	7.23	25.24	**26.38**	**46.46**
ResDAVEnet-L4	Places A+I	42.76	15.02	36.38	32.43	55.45
FHVAE	Aurora4 (Clean)	71.98	4.75	72.54	50.57	86.15
FHVAE	Aurora4 (All)	24.41	5.01	16.42	20.29	36.33

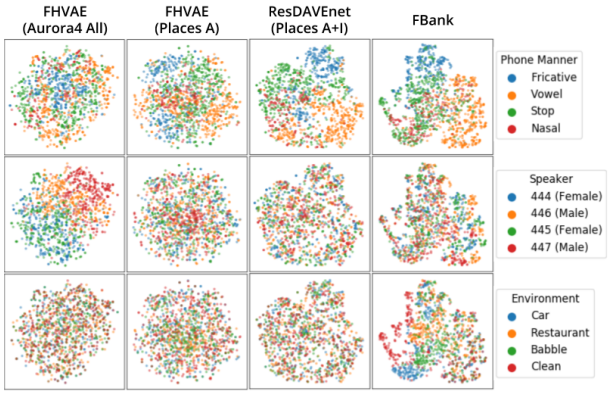

Figure 2: *Frame-level t-SNE projections for four different acoustic representations, color coded for phonetic manner class, speaker identity, and noise/environment type. Visually, the ResDAVEnet features encode the least amount of speaker and environment information.*

t-SNE [52] comparing ResDAVEnet, FHVAE (Places A / Aurora4 All), and the baseline FBank feature. It can be observed from the first row that all three features contain phonetic information, as different phonetic manners are separated in each space. On the other hand, the project features of ResDAVEnet and FHVAE (Aurora4 All) are visually more environment invariant than those from the other two (for the FHVAE trained on PlacesAudCap, green and orange dots concentrate more at the center than red and blue dots). Such visualization correlates well with the performance of the various feature types in Tables 1.

To conclude, we learn that (1) despite being trained with exactly the same process, inductive bias introduced to model architectures (i.e., DAVEnet versus ResDAVEnet) still affects the properties of learned representations, (2) feature extractors distilled from ResDAVEnet models clearly preserve phonetic information while improving invariance to nuisance factors, and most importantly, (3) it achieves better cross-dataset transferrability compared to FHVAE and FBank features.

5.3. Correlation with Source Task Performance

Finally, we study how the performance of the grounding task affects the transfer learning performance, conditioning on the same neural network architecture for the source task. We create two proper subsets of 200k and 80k paired image/audio captions, and train one ResDAVEnet model on each subset. R@10 of the retrieval task for the models trained with 80k, 200k, and 400k (original) are 0.343, 0.582, and 0.720, respectively.

Table 2: *TIMIT test WERs by gender for different ASR features.*

ASR Feature	FE Train Set	Test WER (%) Male	Female
FBank	N/A	**20.39**	31.15
FHVAE	Places A	25.35	33.22
DAVEnet-L1	Places A+I	20.58	30.62
DAVEnet-L2	Places A+I	21.94	32.57
DAVEnet-L3	Places A+I	28.64	32.74
ResDAVEnet-L1	Places A+I	21.48	30.74
ResDAVEnet-L2	Places A+I	22.28	**27.40**
ResDAVEnet-L3	Places A+I	27.26	29.31
ResDAVEnet-L4	Places A+I	38.60	42.07
FHVAE	TIMIT (All)	22.00	26.20

Table 3: *Aurora-4 average test WERs for using features extracted from ResDAVEnet trained on different sizes.*

FE Train Set	L1	L2	L3	L4
Places A+I (80k)	**41.69**	39.52	43.42	51.45
Places A+I (200k)	43.46	37.50	37.85	44.18
Places A+I (400k)	44.03	**33.11**	**33.16**	**42.76**

Results are shown in Table 3. Except for the first layer, we can observe that WER decreases as the amount of source task training data increases. In fact, except for the out-of-domain conditions of the first layer, all layers improve in all conditions (full results not shown due to space limit). Discovery of such positive correlation affirms the relatedness of the two tasks and encourages collection of a larger dataset for building a general feature extractor based on semantic grounding tasks.

6. Concluding Discussion and Future Work

In this paper, we present a successful example of transfer learning from a weakly supervised semantic grounding task to robust ASR. We achieve cross-dataset transferability, which is an important milestone toward building a generalized feature extractor to be used in many tasks and domains like BERT. In addition, along with the analysis in [27, 25], this work sheds light on using semantic level supervision to learn the compositional structure of a language. For future work, we would like to study methods for leveraging target task data, possibly through semi-supervised training or adaptation, in order to bridge the gap to FHVAE trained on those data. Furthermore, unlike FHVAE, it is unclear at which layer a ResDAVEnet model learns to maximally remove domain information. We would also like to explicitly force such disentanglement to occur at certain layers, which can possibly improve both the grounding performance and the robustness of distilled features.

7. References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep Speech 2: End-to-end speech recognition in English and Mandarin," in *ICML*, 2016.

[2] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *ICASSP*, 2018.

[3] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013.

[4] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2015.

[5] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmen-

tation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," 2017.

[7] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *ASRU*, 2017.

[8] W.-N. Hsu, H. Tang, and J. Glass, "Unsupervised adaptation with interpretable disentangled representations for distant conversational speech recognition," in *Interspeech*, 2018.

[9] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "A multi-discriminator CycleGAN for unsupervised non-parallel speech domain adaptation," in *Interspeech*, 2018.

[10] S. Sun, C.-F. Yeh, M. Ostendorf, M.-Y. Hwang, and L. Xie, "Training augmentation with adversarial examples for robust speech recognition," *arXiv preprint arXiv:1806.02782*, 2018.

[11] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE transactions on speech and audio processing*, 1994.

[12] B. E. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *ICASSP*, 1997.

[13] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2016.

[14] J. Fredes, J. Novoa, S. King, R. M. Stern, and N. B. Yoma, "Locally normalized filter banks applied to deep neural-network-based robust speech recognition," *IEEE Signal Processing Letters*, 2017.

[15] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, 2017.

[16] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *NIPS*, 2017.

[17] W.-N. Hsu and J. Glass, "Extracting domain invariant features by unsupervised learning for robust automatic speech recognition," in *ICASSP*, 2018.

[18] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," in *Interspeech*, 2019.

[19] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, 2010.

[20] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *NIPS*, 2016.

[21] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *ECCV*, 2018.

[22] D. Harwath and J. R. Glass, "Learning word-like units from joint audio-visual analysis," in *ACL*, 2017.

[23] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*, 2015.

[24] A. Jansen, M. Plakal, R. Pandya, D. P. Ellis, S. Hershey, J. Liu, R. C. Moore, and R. A. Saurous, "Unsupervised learning of semantic audio representations," in *ICASSP*, 2018.

[25] D. Harwath and J. Glass, "Towards visually grounded sub-word speech unit discovery," in *ICASSP*, 2019.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385.

[27] J. Drexler and J. Glass, "Analysis of audio-visual features for unsupervised speech recognition," in *Grounded Language Understanding Workshop*, 2017.

[28] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014.

[29] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," in *ICLR*, 2015.

[30] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large scale hierarchical image database," in *CVPR*, 2009.

[31] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *CVPR Workshop*, 2014.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[33] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.

[34] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.

[35] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *EMNLP*, 2014.

[36] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *NAACL*, 2018.

[37] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, 2018.

[38] R. A. Yeh, M. N. Do, and A. G. Schwing, "Unsupervised textual grounding: Linking words to image concepts," in *CVPR*, 2018.

[39] T. Gupta, K. Shih, S. Singh, and D. Hoiem, "Aligned image-word representations improve inductive transfer across vision-language tasks," in *ICCV*, 2017.

[40] E. Chuangsuwanich, Y. Zhang, and J. Glass, "Multilingual data selection for training stacked bottleneck features," in *ICASSP*, 2013.

[41] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE/ACM Trans. Audio, Speech & Language Processing*, 2019.

[42] G. Chrupala, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *ACL*, 2017.

[43] A. Alishahi, M. Barking, and G. Chrupala, "Encoding of phonology in a recurrent neural model of grounded speech," in *CoNLL*, 2017.

[44] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech communication*, 1990.

[45] D. Pearce and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," *Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep*, 2002.

[46] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[47] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[48] F. Seide and A. Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *KDD*, 2016.

[49] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech*, 2014.

[50] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNs for distant speech recognition," in *ICASSP*, 2016.

[51] W.-N. Hsu and J. Glass, "Scalable factorized hierarchical variational autoencoder training," in *Interspeech*, 2018.

[52] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *JMLR*, 2008.