INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A DNN-HMM-DNN Hybrid Model for Discovering Word-like Units from Spoken Captions and Image Regions

Liming Wang[1], *Mark Hasegawa-Johnson*[2]

[1]Department of Electrical and Computer Engineering, University of Illinois, Urbana Champaign
[2]Beckman Institute, University of Illinois, Urbana Champaign

lwang114@illinois.edu, jhasegaw@illinois.com

Abstract

Discovering word-like units without textual transcriptions is an important step in low-resource speech technology. In this work, we demonstrate a model inspired by statistical machine translation and hidden Markov model/deep neural network (HMM-DNN) hybrid systems. Our learning algorithm is capable of discovering the visual and acoustic correlates of K distinct words in an unknown language by simultaneously learning the mapping from image regions to concepts (the first DNN), the mapping from acoustic feature vectors to phones (the second DNN), and the optimum alignment between the two (the HMM). In the simulated low-resource setting using MSCOCO and SpeechCOCO datasets, our model achieves 62.4 % alignment accuracy and outperforms the audio-only segmental embedded GMM approach on standard word discovery evaluation metrics.

Index Terms: unsupervised spoken word discovery, multimodal learning, language acquisition, machine translation

1. Introduction

Multimodal word discovery is the problem of discovering words, in a previously unknown language, using a database in which spoken utterances are matched to semantically related content in some other modality. In a previous paper [1], we reformulated the multimodal word discovery problem as an analog of statistical machine translation (SMT), and we proposed learning an alignment between a sequence of phones (labeled in the international phonetic alphabet) and a set of image concepts (each matched to an underlying bounding box in the image). The SMT approach has some advantages over the matched embedding approach of [2], e.g., it performs word discovery simultaneous with the refinement of the discovered word models, so that the resulting models of concept-to-word alignment (the alignment model) and of word-to-phone generation (the likelihood model) are simultaneously optimal. In [1], however, the phones and image concepts are both discrete, therefore the algorithm is only useful if the image and audio are pre-processed by an object detection algorithm and a phone recognizer, respectively. This paper extends the model of [1] (significantly) by training deep neural posterior probability models, capable of learning simultaneously optimal alignment and posterior models for real-valued image and audio feature vectors. The alignment model is implemented as a sort of HMM, in which the discrete state (image concept) and discrete observable (phone) are each related by DNNs to real-valued feature vectors (image and speech, respectively), hence the system could be called a DNN-HMM-DNN hybrid model.

2. Related Works

Human infants develop associations between acoustic stimuli and visible objects through a process of language acquisition, e.g., Skinner proposed that all language is learned by multimodal association [3]. The task of teaching computers to learn language by multimodal association was proposed in three simultaneous research efforts in the late 1990s [4, 5, 6]. All three projects used mobile or partially mobile robots, and in all three projects, the movement of the robot was used to align audio and visual stimuli: researchers spoke the name of objects within the robot's visual field. Robots learned to associate the audio and visual stimulus using a replay memory [4], a hierarchy of HMMs [7], or a graphical model [6].

In [8], crowd workers were hired to read the captions in an image captioning corpus [9], creating a standardized corpus for multimodal word discovery. In the first papers using this corpus, word discovery was framed as an information retrieval problem: a cosine distance between learned embeddings was used to retrieve images from audio, or audio from images. The pre-trained image embedding [10] provided supervision for training of the audio embeddings, either with [8] or without [11] ground truth word boundaries. In [2], it was demonstrated that the same embeddings could be used to automatically discover word boundary times in the audio, and object bounding boxes in the image, by exhaustively searching over a grid of audio and image segments. The exhaustive search was replaced in [12] by a more efficient convolutional time alignment, in which peaks in similarity between the image and audio convolutional networks were taken to indicate discovered image concepts and audio words, respectively. Convolutional multimodal time alignment is able to automatically discover word alignments between Hindi and English [13], and to discover phone-like units in speech [14].

Multimodal word discovery is closely related to the problem of unsupervised acoustic unit discovery, which tries to cluster syllable-like units from raw audio. Unsupervised learning of acoustic units can often be decomposed into two problems [15]: segmentation divides the audio into variable-length segments, then clustering of the segments creates an inventory of discovered acoustic units. Some models detect syllable boundaries using an algorithm inspired by human speech perception and cluster by searching for recurrent patterns in the speech [16]. Other models attempted to solve the segmentation and clustering problem jointly using a nonparametric Bayesian network that models each subword as a hidden Markov model (HMM) [15] or Gaussian mixture model (GMM) [17]. To improve the purity of the discovered clusters, several prior distributions are used, such as Pitman-Yor process [18], symmetric Dirichlet prior [15, 17], stick-breaking process [19] and prior based on ℓ_1 norm of the observation probabilities [20].

http://dx.doi.org/10.21437/Interspeech.2020-1148

3. Problem Formulation

Suppose a language learner tries to learn a set of unknown visually salient words from a spoken image caption with acoustic features $\mathbf{x} = [x_1, x_2, \cdots, x_T] \in \mathbb{R}^{D_x \times T}$ and a set of image regions with features $\mathbf{y} := [y_1, y_2, \cdots, y_L] \in \mathbb{R}^{D_y \times L}$. Assume that each image region y_i depicts a single, discrete *image concept* z_i, and that the image concepts $\mathbf{z} := [z_1, \ldots, z_L]$ are independent, identically distributed drawn from a finite collection of concepts $\mathcal{C} = \{1, \cdots, K\}$, where K is the maximum number of distinct concepts the learner can learn in one phase of acquisition. Assume that each acoustic feature x_t represents a single acoustic phone unit ϕ_t drawn from the phone set $\Phi = \{1, \cdots, V\}$ and aligns to at most one image concept z_{i_t}. Suppose the alignments $\mathbf{i} := [i_1, \cdots, i_T]$, $i_t \in \{1, \ldots, L\}$ are generated by a Markov chain $p(i_t|i_{t-1}, L)$, and the phone features $\phi := [\phi_1, \cdots, \phi_T]$ are drawn independently given their aligned image regions, then the conditional likelihood $p(\mathbf{x}|\mathbf{y})$ can be written as:

$$p(\mathbf{x}|\mathbf{y}) = \sum_{\mathbf{z},\mathbf{i},\phi} p(\mathbf{z}|\mathbf{y}) p(\mathbf{i}, \phi, \mathbf{x}|\mathbf{z}, L) \tag{1}$$

$$p(\mathbf{z}|\mathbf{y}) = \prod_{i=1}^{L} p(z_i|y_i) \tag{2}$$

$$p(\mathbf{i}, \phi, \mathbf{x}|\mathbf{z}, L) = \prod_{t=1}^{T} p(i_t|i_{t-1}, L) \sum_{\phi_t \in \Phi} p(\phi_t|z_{i_t}) p(x_t|\phi_t) \tag{3}$$

Further, the *unimodal concept posterior* probability of the image concepts, $p(z_i = k|y_i) =: \pi_{ik} =: \pi_k(y_i)$ is assumed to be a softmax distribution of some kernel functions $g_k, k \in \mathcal{C}$ referred later as the *visual kernels*:

$$\pi_{ik} = \frac{\exp(g_k(y_i))}{\sum_{k'=1}^{K} \exp(g_{k'}(y_i))} \tag{4}$$

To allow a more general class of observation probability distributions for the acoustic features, we approximate $p(x_t|\phi_t)$ with the *unimodal phone posterior* probability $p(\phi_t = \phi|x_t) =: b_{t\phi} =: b_\phi(x_t)$ by assuming uniform prior on the phones and the segment features:

$$b_{t\phi} := \frac{\exp(h_\phi(x_t))}{\sum_{\phi' \in \Phi} \exp(h_{\phi'}(x_t))} \tag{5}$$

$$p(x_t|\phi_t) = \frac{p(x_t) p(\phi_t|x_t)}{p(\phi_t)} \tag{6}$$

$$\propto p(\phi_t|x_t) = b_{t\phi_t}, \tag{7}$$

where $h_\phi(\cdot), \phi \in \Phi$ are some kernel functions called the *acoustic kernel* similar to the visual kernel.

The goal of our model is to infer the hidden concepts $\mathbf{z}$, the hidden phone clusters ϕ and the hidden alignments $\mathbf{i}$ between each image region and subsets of phone segments that describe that image region, with a maximum *a posteriori* rule:

$$\mathbf{z}^*, \phi^*, \mathbf{i}^* = \arg\max_{\mathbf{z},\phi,\mathbf{i}} p(\mathbf{z}, \phi, \mathbf{i}|\mathbf{x}, \mathbf{y}). \tag{8}$$

4. Model Description

Our model consists of three main components: An HMM alignment model, a concept posterior model learned by a DNN, and a phone posterior model learned by another DNN.

If Eq. (2) were a product over time ($\prod_{t=1}^{T} p(z_{i_t}|y_{i_t})$) rather than over image regions ($\prod_{i=1}^{L} p(z_i|y_i)$), then Eq. (1) would be a time-synchronized HMM, and could be solved using the forward-backward algorithm. However, that is not the case since we assume each image region represents a unique concept.

The expectation maximization (EM) algorithm is used to optimize Eq. (1) with respect to the initial probability $p(i_t|L)$ and transition alignment probabilities $p(i_t|i_{t-1}, L)$ and the concept-to-phone probabilities $p(\phi_t|z_i)$. While the full latent variable space has exponential complexity ($\mathcal{O}\left\{K^L\right\}$), it turns out that under the Markov assumption, we only need to evaluate a small subset of the latent posterior probabilities to update the parameters of the alignment model using the *partial state* variables $s_t := (i_t, z_{i_t}), t = 1, \cdots, T$. The forward-backward algorithm over s_t has forward probabilities $\alpha_t(i, k) := p(\mathbf{x}_{1:t}, z_i = k, i_t = i|\mathbf{y})$ and backward probabilities $\beta_t(i, k) := p(\mathbf{x}_{t+1:T}|z_i = k, i_t = i, \mathbf{y})$ that contain the concept posterior of exactly one image region at a time, e.g.,

$$\alpha_t(j, l) = \sum_{i=1}^{L} \sum_{k=1}^{K} \alpha_{t-1}(i, k) a_{ij} p(l|i, j, k, y_j) p(x_t|k), \tag{9}$$

where $p(l|i, j, k, y_i) = \mathbb{1}[k = l]$ if $i = j$, else π_{jl}; $a_{ij} = p(i_t = j|i_{t-1} = i, L)$, and $p(x_t|k) = \sum_{\phi_t} p(\phi_t|z_{i_t} = k) b_{t\phi}$.

For the unimodal concept and phone posteriors, gradient ascent with respect to the log conditional likelihood $L_{MLE} = \log p(\mathbf{x}|\mathbf{y})$ is used to update their parameters. The gradients propagating to the visual kernels and the acoustic kernels can be shown respectively to be:

$$\frac{\partial L_{MLE}}{\partial g_{ik}} = \frac{1}{p(\mathbf{x}|\mathbf{y})} \sum_{j=1}^{K} \frac{p(\mathbf{x}, z_i = j|\mathbf{y})}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial g_{ik}} \tag{10}$$

$$= p(z_i = k|\mathbf{x}, \mathbf{y}) - \pi_{ik} =: \Delta_{ik}^{image} \tag{11}$$

$$\frac{\partial L_{MLE}}{\partial h_{t\phi}} = p(\phi_t = \phi|\mathbf{x}, \mathbf{y}) - b_{t\phi} =: \Delta_{t\phi}^{phone}. \tag{12}$$

Intuitively, the gradients for modalities are positive in the direction that moves the concept and phone unimodal posteriors (π_{ik} and $b_{t\phi}$) closer to the multimodal posterior probabilities of concept k and phone ϕ, respectively. For example, when $g_k(y)$ (similarly $h_\phi(x)$) is a scaled Euclidean kernel, $g_k(y) = -\|y - \mu_k^{im}\|_2^2 / 2\sigma^2$, then Eq. (4) becomes a scaled Gaussian, whose mean is re-estimated as

$$\mu_k^{image} = \frac{\sum_{i=1}^{L} \Delta_{ik}^{image} y_i}{\sum_{i=1}^{L} \Delta_{ik}^{image}} \tag{13}$$

In general, the parameters θ of the kernel functions satisfy:

$$\frac{\partial L_{MLE}}{\partial \theta_k^{image}} = \sum_{i=1}^{L} \sum_{k'=1}^{K} \Delta_{ik'}^{image} \frac{\partial g_{k'}(y_i)}{\partial \theta_k^{image}}, \tag{14}$$

where the gradient step Δ_{ik}^{image} is computed from $p(z_i = k|\mathbf{x}, \mathbf{y})$, which can be computed in $O(L^3 KT)$ time using a *biased* version of the forward probability. Define the biased forward probability to be $\alpha_t(i|j, k) := p(\mathbf{x}_{1:t}, i_t = i|z_j = k, \mathbf{y})$, for $i = 1, \ldots, L, k = 1, \ldots, K$, then:

$$p(z_i = k|\mathbf{x}, \mathbf{y}) = \frac{\pi_{ik} \sum_{j=1}^{L} \alpha_T(j|i, k)}{\sum_{k=1}^{K} \pi_{ik} \sum_{j=1}^{L} \alpha_T(j|i, k)}. \tag{15}$$

Gradient update of θ_{ϕ}^{phone} is similar, except that the multimodal phone posterior $p(\phi_t = \phi|\mathbf{x}, \mathbf{y})$ is a direct byproduct of the unbiased EM algorithm:

$$p(\phi_t|\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{L} \alpha_t(i, z) p_t(\phi_t|z) \beta_t(i, z)}{\sum_{i=1}^{L} \sum_{z=1}^{K} \alpha_t(i, z) p_t(\phi_t|z) \beta_t(i, z)}, \quad (16)$$

where $p_t(\phi_t|z) = \frac{p(\phi_t|z)p(x_t|\phi_t)}{\sum_{\phi_t \in \Phi} p(\phi_t|z)p(x_t|\phi_t)}$.

Since our model has an exponential state space, simultaneously decoding $\mathbf{z}^*$, $\mathbf{i}^*$ and ϕ is intractable, but a reasonable approximation can be implemented using a type of coordinate descent. Using a Viterbi algorithm similar in form to Eq. (9), $\mathbf{i}^* = \operatorname{argmax} p(\mathbf{i}|\mathbf{x}, \mathbf{y})$ can be optimized with an implicit average over all possible $\mathbf{z}$, then $\mathbf{z}^*$ can be optimized element-wise assuming a known $\mathbf{i}^*$:

$$\mathbf{z}^* \approx \operatorname*{argmax}_{\mathbf{z}} p(\mathbf{z}|\mathbf{i}^*, \mathbf{x}, \mathbf{y}) \propto \prod_{i=1}^{L} \prod_{t:i_t=i} p(x_t|z_i) \quad (17)$$

5. Experimental Setup

In order to simulate a low-resource setting, we constructed a small synthetic dataset (called hereafter MSCOCO 2k) by sampling image regions and matching spoken concept names from the MSCOCO [21] and SpeechCOCO [22] validation sets, respectively. Of the 80 object classes labeled in MSCOCO, we selected the 65 whose class names appear in the original spoken captions. For each such class, we randomly selected MSCOCO images containing regions labeled with that class, with a maximum of 200 matched images per class. Image feature vectors were extracted from each of the selected image regions, and concatenated to form 2541 simulated images, each of which contained exactly five image regions. One spoken instance of each class name was extracted from the original SpeechCOCO caption, and the five speech segments were concatenated to form a simulated audio caption.

Acoustic feature vectors were computed once per phone, and concatenated to form an acoustic feature sequence $\mathbf{x}$ whose duration T is equal to the total number of phones in the audio file. We experimented with three different phone-level features for the clustering model. The "re-sample" approach was similar to [17] and was created by re-sampling the Mel-Frequency Cepstral Coefficients ([23]) features with 25 ms window, 10 ms skip step and 14 cepstral coefficients for each ground truth phone segments to a 140-dimensional embedding. The "CTC mean" feature was based on a two-layer Long Short-term Memory (LSTM) [24], with 100-dimensional hidden layers, pre-trained using the Connectionist Temporal Classification (CTC) training criterion [25] based on ground truth phone label sequences in the SpeechCOCO training corpus (which does not overlap our MSCOCO 2k corpus). This phone recognizer was trained using stochastic gradient descent with greedy layer-wise pre-training and a learning rate of 10^{-5}. Inputs were MFCC. The phone error rate on MSCOCO2k is 20 %. The "CTC mean" feature vector is then computed as the average, within each ground truth phone segment, of the 100-dimensional hidden node activations in the last LSTM layer of the recognizer. Finally, the "force-aligned phones" feature vector was computed using the predicted phone-label outputs of the same phone recognizer.

Image feature vectors were extracted from ground truth image object regions using the residual net image classifier with 34 layers [26]; image features are the embedding vectors from the penultimate layer of the classifier. Both the phone recognizer and the image classifier used in this work are implemented in Pytorch [27] (https://github.com/lwang114/MultimodalWordDiscovery).

Table 1: *Concept-to-phone word discovery results on MSCOCO2k (in %)*

	Segmental (HMM)	Adaptor Grammar [18]
Alignment Acc	97.0	-
Grouping F1	96.1	-
Boundary F1	90.0	**93.1**
Token/Type F1	77.1/24.0	**85.1/62.0**

The unimodal posterior models of the DNN-HMM-DNN use the Gaussian kernel function defined in Eq. (13) with the width $\sigma^2 = 1$ for all experiments. The alignment initial probabilities, the alignment transition probabilities and the phone probabilities are initialized uniformly over their supports. The gradient ascent algorithm for the clustering models uses the natural gradients Eq. (15) over the entire dataset with a learning rate of 0.1.

To study the role played by different components of our image-to-audio system, we conducted two ablation studies: (1) the model is trained to map phones to ground truth image labels (*concept-to-phone* discovery), (2) the model is trained to map audio feature vectors to ground truth image labels (*concept-to-audio* discovery). For concept-to-phone discovery, we used the adaptor grammar as our baseline [18]. For concept-to-audio and image-to-audio discovery, we used the segmental GMM [17] as the baseline. We used the code provided by [17] and initialized the landmark boundaries to be the ground truth phone boundaries. This pre-segmentation ensures the same level of supervision for all models in the acoustic modality, therefore makes the results more reflective of the extent to which adding the visual modality benefits the word discovery system. We also compared our results to the adaptor grammar trained on force-aligned phones and another image-to-audio word discovery system, DAVENet [13] trained on the much larger Places 400k dataset [28] with 101 concept classes, which despite its size, has a similar number of concepts to the 65 concept classes we have.

6. Results

Table 1 presents results of the concept-to-phone ablation study, compared to the word discovery performance of an adaptor grammar [18]. The adaptor grammar is an algorithm that forms words by clustering phone sequences so that lexical usage frequencies follow a Pitman-Yor distribution; it has no access to the image concept labels. Metrics reported in the table are from the Zero Resource Speech Challenge (ZSRC) [29] computed using tools provided by the authors; they include alignment accuracy, grouping F1, boundary F1, and token/type F1. An alignment is a link between a phone and a concept: in each spoken caption, each phone is correctly aligned with exactly one concept. A grouping is a link between two phones: in each spoken caption, phones in the same concept are correctly grouped. A boundary is correct if it divides consecutive phones that belong to different concepts, and incorrect otherwise. Type F1 measures recall and precision in the detection of unique phone strings as candidate words; Token F1 is the same measure, but weights each phone string by its frequency. Our model achieves 97% alignment accuracy but perform worse in the boundary

Table 2: *Concept-to-audio word discovery results on MSCOCO 2k (in %)*

A - Alignment Accuracy GF1 - Grouping F1
BF1 - Boundary F1 TF1 - Token/Type F1

	Re-sample	CTC Mean	Force-Aligned Phones	Segmental GMM
A	72.9	75.2	**85.3**	-
GF1	54.9	61.6	**68.6**	25.2
BF1	43.2	47.2	**55.8**	43.7
TF1	17.1/4.3	12.2/1.2	**32.7/5.2**	11.6/3.3

and token/type F1 score than the adaptor grammar, suggesting that the assignment of tokens to types is improved by imposing a Pitman-Yor prior. Indeed, the reason for the large discrepancy between the token and type F1 score for our models is not because of an imbalanced dataset (the class distribution of MSCOCO2k is uniform), but because of the within-class variability of the phone strings discovered by our models. Due to a lack of word-level contextual information, we found that the incorrect 3% of the alignment creates 700 additional incorrectly discovered word classes, which drag down the type precision of our model.

Concept-to-audio word discovery results are shown in Table 2. In this case, we compare our models with the audio-only segmental GMM [17] approach. Consistent with the concept-to-phone word discovery results, the multimodal models consistently perform better in the grouping F1 scores. Further, the multimodal model also generally performs better in the boundary and token F1 scores than the audio-only segmental GMM baseline, suggesting that the visual modality increases the robustness of the model to acoustic variability. Among different multimodal models, the discrete force-aligned phone features perform better than the continuous re-sampled and CTC-mean features, suggesting that the continuous features tend to have a higher signal-to-noise ratio than the discrete labels for our dataset. Comparing the fully unsupervised re-sampled acoustic embedding feature with the CTC-mean feature, we observe that the re-sampled feature has better token and type F1, worse grouping F1, and comparable alignment and boundary scores. This result might be interpreted to mean that the CTC pre-training improves its ability to identify grouping, but provides no information that improves alignment or token/type discovery: our proposed multimodal alignment is apparently sufficient supervision for token/type discovery, even without using pre-trained acoustic features.

Finally, the end-to-end word discovery result using image and acoustic features are shown in Table 3. This table includes three new measures: cluster purity (the percentage of acoustic feature vectors assigned to each discovered word type whose ground-truth label matches the majority), coverage (the percentage of ground truth phones that are chosen by the system for alignment to any acoustic feature vector), and average length (of discovered word tokens). Baselines are the adaptor grammar, which sees only audio features, and DAVENet [13], which sees both audio and images. In this case, for the "Res34+Forced-Aligned" model, we no longer require ground truth phone boundaries; instead, we pass the whole audio caption to our CTC recognizer to obtain the force-aligned phone transcripts. For the clustering performance, we can see that our model generates slightly lower-quality clusters but achieves better coverage than DAVENet. Our model generally has trouble learning good hidden representations of the image concepts even when it is able to learn the alignment primarily due to our decoding method, which decodes the alignment by averaging out all the image concepts. As a result, each individual concept variable correlates less well with the true concept of the features. Again the discrete feature leads to better results than the continuous feature. The drop in alignment accuracy from discrete to continuous features is comparable to that observed in Table 2, suggesting that the effects of acoustic feature variability (between these two columns) and of visual feature variability (between these two tables) are approximately independent.

Table 3: *Image-to-audio word discovery results on MSCOCO 2k (in %)*

GF1 - Grouping F1 BF1 - Boundary F1
TF1 - Token/Type F1

	Res 34 + Force-Aligned	Res 34 + CTC-Mean	Adaptor Grammar	DAVENet (Places 400k) [13]
Accuracy	**62.4**	53.3	-	-
Purity	46.1	46.1	-	**53**
GF1	**41.1**	32.0	-	-
BF1	**48.0**	47.3	47.5	-
TF1	17/4.3	7.7/3.1	**28.2/11.9**	-
Coverage	87	**100**	-	45
Avg. Len (GT=5.56)	**6.34**	6.35	4.41	-

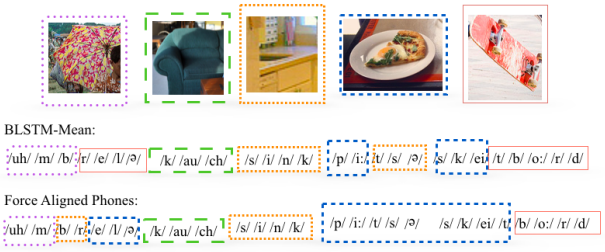

Figure 1: *An example of the image-to-audio word discovery result. The inputs of the algorithm are acoustic phone segments and image regions. The ground truth phone labels are not available during training and only shown for clarity. The phone segment and image region with matching color frames are aligned by the models.*

An example of the word discovery result is shown on Fig. 1. Our models are able to discover words such as "sink" and "couch". But due to a lack of word-level context information, our model tends to oversegment longer words such as "umbrella" and "skateboard" and sometimes confuses between two concepts that share phones such as "um**br**ella" and "skate**b**oa**r**d".

7. Conclusion

In this work, we presented a DNN-HMM-DNN hybrid model to learn word units from audio and semantically related images. Our model can be optimized end-to-end efficiently using an exact EM algorithm and achieves better word discovery performance than the audio-only approaches.

8. References

[1] L. Wang and M. Hasegawa-Johnson, "Multimodal word discovery and retrieval with phone sequence and image concepts," in *Interspeech*, 2019.

[2] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," in 55^{th} *Annual Meeting of the Association for Computational Linguistics*, 2017.

[3] B. F. Skinner, *Verbal Behavior*. New York: Copley Publishing Group, 1957, 1992.

[4] D. Roy, "A computational model of word learning from multimodal sensory input," in *Proceedings of the international conference of cognitive modeling*, Groningen, The Netherlands, 2000, pp. 1–8.

[5] S. E. Levinson, Q. Liu, C. Dodsworth, R.-S. Lin, W. Zhu, and M. Kleffner, "The role of sensorimotor function, associative memory and reinforcement learning in automatic acquisition of spoken language by an autonomous robot," in *Joint NSF DARPA Workshop on Development and Learning*, East Lansing, MI, 2000.

[6] N. Iwahashi, "Language acquisition through a human–robot interface by combining speech, visual, and behavioral information," *Information Sciences*, vol. 156, no. 1-2, pp. 109–121, 2003.

[7] S. E. Levinson, *Mathematical Models of Spoken Language*. John Wiley and Sons, 2005.

[8] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," *Automatic Speech Recognition and Understanding*, 2015.

[9] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[11] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Neural Information Processing Systems*, 2016.

[12] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 27, pp. 89–98, 2019.

[13] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[14] D. Harwath and J. Glass, "Towards visually grounded sub-word speech unit discovery," in *International Conference on Acoustics, Speech and Signal Processing*, 2019.

[15] C. 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*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 40–49.

[16] O. J. Raŝañen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Interspeech*, 2015.

[17] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 24, pp. 669–679, 2016.

[18] M. Johnson, T. Griffiths, and S. Goldwater, "Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models," in *Neural Information Processing Systems*, 2007.

[19] L. Ondel, P. Godard, L. Besacier, E. Larsen, M. Hasegawa-Johnson, O. Scharenborg, E. Dupoux, L. Burget, F. Yvon, and S. Khudanpur, "Bayesian models for unit discovery on a very low resource language," in *Proc. ICASSP*, 2018.

[20] S. Bharadwaj, M. Hasegawa-Johnson, J. Ajmera, O. Deshmukh, and A. Verma, "Sparse hidden Markov models for purer clusters," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[21] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, 2014.

[22] W. Havard, L. Besacier, and O. Rosec, "Speech-COCO: 600k visually grounded spoken captions aligned to MSCOCO data set," in *GLU 2017 International Workshop on Grounding Language Understanding*, 2017.

[23] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[25] A. Graves, S. Férnandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, T. Gregory, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[28] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Neural Information Processing Systems (NeurIPS)*, 2014, p. 487–495.

[29] E. Dunbar, X. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," *CoRR*, vol. abs/1712.04313, 2017. [Online]. Available: http://arxiv.org/abs/1712.04313