UNSUPERVISED AUDIO-CAPTION ALIGNING LEARNS CORRESPONDENCES BETWEEN INDIVIDUAL SOUND EVENTS AND TEXTUAL PHRASES

Huang Xie, Okko Räsänen, Konstantinos Drossos, Tuomas Virtanen

Unit of Computing Sciences, Tampere University, Finland

ABSTRACT

We investigate unsupervised learning of correspondences between sound events and textual phrases through aligning audio clips with textual captions describing the content of a whole audio clip. We align originally unaligned and unannotated audio clips and their captions by scoring the similarities between audio frames and words, as encoded by modality-specific encoders and using a ranking-loss criterion to optimize the model. After training, we obtain clip-caption similarity by averaging frame-word similarities and estimate event-phrase correspondences by calculating frame-phrase similarities. We evaluate the method with two cross-modal tasks: audio-caption retrieval, and phrase-based sound event detection (SED). Experimental results show that the proposed method can globally associate audio clips with captions as well as locally learn correspondences between individual sound events and textual phrases in an unsupervised manner.

Index Terms— Cross-modal learning, audio, caption, sound event, unsupervised learning

1. INTRODUCTION

Cross-modal learning, which aims at processing and relating information across multimodal data (e.g., audio, image and text), has received increasing attention recently. In this paper, we focus on cross-modal learning between audio and natural language descriptions, i.e., audio-text cross-modal learning. Generally, natural language allows the description of acoustic information and the versatile modeling of sound relationships in ways that are understandable by humans. Automated interpretation of audio data with natural language has great potential in real-world applications, such as audio retrieval, acoustic monitoring, and human-computer interaction.

Recent audio research that deals with cross-modal learning across audio and text modalities includes audio-text retrieval [1], automated audio captioning [2, 3], and audio question answering [4]. Audio-text retrieval concerns matching of audio and text pairs. Elizalde et al. [1] associated audio with text by jointly learning representations of audio and text with a siamese network. In automated audio captioning [2, 3], captions were automatically generated for audio clips to summarize sound events contained in them. Audio question answering investigates acoustic reasoning through answering textual questions pertaining to audio clips. The pioneering work [4] predicted answers across predefined values by learning joint representations of audio and questions.

The aforementioned studies [1, 2, 3, 4] mainly focus on information fusion across audio and text modalities, for example, learning joint audio-text representations. However, the relationships between audio and text elements are rarely investigated. For example, audio clips consists of sound events, which are described by textual phrases in human-annotated captions [5, 6]. It is not fully understood what information the above cross-modal learning methods learn when matching whole audio clips and their captions.

Inferring the latent relationships between audio and text elements is a key to automated interpretation of audio data using textual descriptions. The problem of finding relationships between elements from different modalities is referred to as cross-modal alignment, which has been widely studied in the language and vision communities [7]. For example, given an image and a written or spoken caption, cross-modal alignment aims to associate the image regions with the caption's words or phrases [8, 9, 10, 11, 12, 13, 14]. These studies have shown the potential for learning element-wise correspondences across different modalities via cross-modal alignment. Particularly, Harwath et al. [12] demonstrate that visual objects and spoken words can be jointly discovered from raw images and speech signals by aligning image pixels with speech waveform in an unsupervised manner. In contrast to supervised alignment, which requires labeled aligned training data, unsupervised alignment can explore cross-modal correspondences without resorting to expensive data annotations.

We propose an unsupervised audio-text aligning method to investigate the correspondences between sound events and textual phrases within clip-caption pairs. Previous work [15] explored event-phrase correspondences with a supervised approach, where timestamps and textual phrases of sound events were provided for clip-caption pairs during model training. In contrast, we utilize audio clips and captions that are temporally unaligned and unannotated, aside from knowing which of them belong together. We obtain clip-caption similarities by aggregating frame-word similarities. The proposed method is evaluated on two cross-modal tasks: audio-caption retrieval and phrase-based SED. Experimental results show that the proposed method can learn global associations between audio clips and captions as well as local correspondences between individual sound events and textual phrases.

2. PROPOSED METHOD

In this section, we present the proposed unsupervised audio-text aligning method for learning event-phrase correspon-

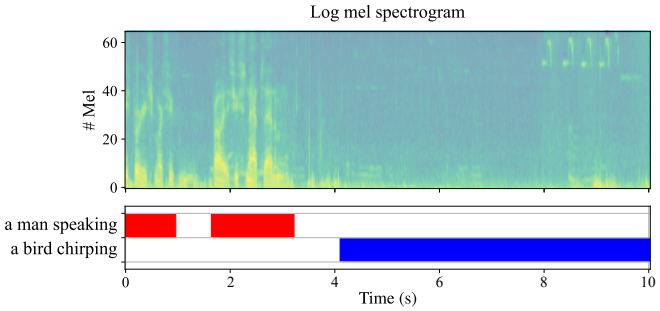

Fig. 1. An example of the temporal presences of two sound events in an audio clip and their corresponding descriptive phrases in the clip's caption. The audio clip is illustrated with its log mel spectrogram, and the temporal activity of each sound event is shown with colored bars.

dences and clip-caption similarities. Figure 1 shows an example clip-caption pair containing two sound events and their corresponding textual phrases which are targeted in this study.

2.1. Audio-Text Aligning Framework

Inspired by previous works [12, 15], we propose an audio-text aligning framework with two input encoders: one for audio, and the other for text, as illustrated in Figure 2. Alignment between audio clip x and its caption y involves processing and relating audio information in x and text information in y, as well as mapping together similarities of their acoustics and lexical semantics. We split x into a sequence of N audio frames $(f_i^x)_{i=1}^N$ and represent y by a sequence of M words $(w_j^y)_{j=1}^M$. A textual phrase p is defined by a sub-sequence $(w_l^y, \ldots, w_m^y)$ of y, which ranges from the l-th to m-th words.

The audio encoder extracts frame-wise acoustic embeddings $(\mathbf{a}_i^x)_{i=1}^N \in \mathbb{R}^{300}$ from x, and the text encoder converts y into a sequence of word-specific semantic embeddings $(\mathbf{b}_j^y)_{j=1}^M \in \mathbb{R}^{300}$. A frame-word alignment matrix $\mathbf{W}_{xy} \in \mathbb{R}^{N\times M}$ of x and y is obtained by scoring the similarity between each audio frame and word pair.

Following [12], we define the similarity of the i-th audio frame f_i^x and the j-th word w_j^y by the dot product of their embeddings $\mathbf{a}_i^x$ and $\mathbf{b}_j^y$

$$sim(f_i^x, w_j^y) = dot(\mathbf{a}_i^x, \mathbf{b}_j^y). \tag{1}$$

To associate audio clips with captions, we rely upon the positive frame-word similarities and discard the negatives. Therefore, the alignment matrix $\mathbf{W}_{xy}$ can be written as

$$\mathbf{W}_{xy} = (s_{ij}(x,y)) \in \mathbb{R}^{N\times M}, \tag{2}$$

where $s_{ij}(x,y) = \max(0, sim(f_i^x, w_j^y))$ is the trimmed similarity of the i-th audio frame f_i^x and the j-th word w_j^y. With $\mathbf{W}_{xy}$, we calculate the global clip-caption similarity $S(x,y)$ between x and y by averaging the matrix elements

$$S(x,y) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} s_{ij}(x,y). \tag{3}$$

We train the audio-text aligning framework by optimizing a ranking-based criterion [16], such that audio clips and captions that belong together are more similar than mismatched clip-caption pairs. Specifically, across a batch of K clip-caption pairs $\{(x_k, y_k)\}_{k=1}^K$, where y_k is a caption pertaining to an audio clip x_k, we randomly select imposter clip $\hat{x}_k$ and imposter caption $\hat{y}_k$ for each clip-caption pair (x_k, y_k). Then, we calculate the widely used sampling-based triplet loss [8, 12, 17, 18, 19]

$$\begin{aligned} loss = \frac{1}{K}\sum_{k=1}^{K}[&\max(0, S(x_k,\hat{y}_k) - S(x_k,y_k) + \eta) \\ &+ \max(0, S(\hat{x}_k,y_k) - S(x_k,y_k) + \eta)], \end{aligned} \tag{4}$$

where η is a margin hyper-parameter. We fix η to one.

2.2. Phrase-based SED

The phrase-based SED refers to detecting sound events within an audio clip based on textual phrases describing them. It can be employed as a proxy task for investigating what the model presented in the previous section learns. Sound events are detected by examining the similarities between audio frames and textual phrases. We obtain frame-phrase similarities by aggregating $\mathbf{W}_{xy}$ along words contained in p. The frame-phrase similarity $T(f_i^x, p)$ is calculated by aggregating the trimmed frame-word similarities $\{s_{il}(x,y), \ldots, s_{im}(x,y)\}$

$$T(f_i^x, p) = aggregation(s_{il}(x,y), \ldots, s_{im}(x,y)). \tag{5}$$

Then, min-max normalization is applied to the frame-phrase similarities $(T(f_i^x, p))_{i=1}^N$ across all the audio frames of one audio clip. For event detection, event e is predicted to be present in f_i^x if $T(f_i^x, p)$ is above a detection threshold γ.

2.3. Audio Encoder

We use a convolutional recurrent neural network (CRNN) [3] as the audio encoder. It consists of five convolution blocks, followed by a bidirectional gated recurrent unit (BiGRU). Each convolution block includes an initial batch normalization, a convolutional layer with padded 3×3 convolutions, and a LeakyReLU activation with a slope of -0.1. After the first, third, and fifth convolution blocks, one L4-Norm subsampling layer is used to reduce the temporal dimension of each block's output by a factor of four. A dropout layer with a rate of 0.3 is placed between the last L4-Norm layer and the BiGRU. Lastly, an up-sampling operation is applied to ensure the final output has the same temporal dimension as the CRNN input.

The CRNN audio encoder takes 64-dimensional log mel-band energies as input. Each audio clip is split into 40 ms Hanning-windowed frames with a hop length of 20 ms. Then, 64 log mel-band coefficients are extracted from each frame. A sequence of 300-dimensional acoustic embeddings are generated for each audio clip. The final acoustic embeddings are normalized with L2 normalization.

2.4. Text Encoder

We utilize Word2Vec (Skip-gram model) [20] as the text encoder. Word2Vec is a two-layer fully-connected neural net-

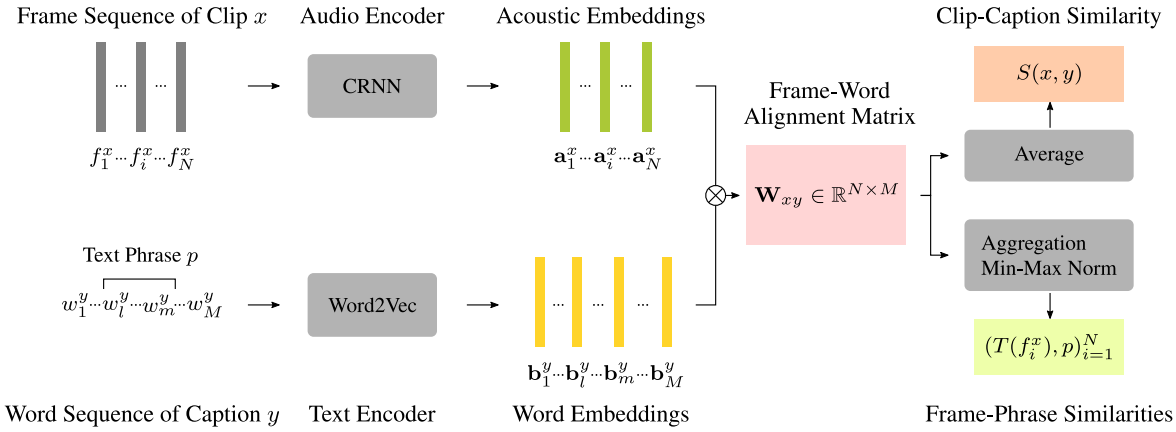

Fig. 2. The proposed audio-text aligning framework.

work, which learns word embeddings that are good at predicting surrounding words in a sentence or a document. For the sake of simplicity, we adopt a publicly available pre-trained Word2Vec [21], which is trained on Google News dataset. It consists of 300-dimensional word embeddings for roughly three million case-sensitive English words and phrases. The Word2Vec text encoder converts textual descriptions into sequences of semantic word embeddings word by word. The final word embeddings are normalized with L2 normalization, and then fixed in our experiments.

3. EXPERIMENTS

In this section, we evaluate the proposed method with the AudioGrounding dataset [15] on two cross-modal tasks: audio-caption retrieval, and phrase-based SED.

3.1. Dataset

The original AudioGrounding dataset [15] consists of 4,590 10-second audio clips, 4,994 descriptive captions, and 13,985 sound event phrases. It is split into three sets: a training set with 4,489 clips, a validation set with 31 clips, and a test set with 70 clips. All audio clips are drawn from YouTube videos, and their corresponding captions are sourced from Audiocaps [5]. One human-annotated descriptive caption is provided for each clip in the training set while five captions are provided for each in the validation and test sets. Sound event phrases are extracted automatically from captions, and human annotators are invited to merge the extracted phrases that correspond to the same sound event and provide the timestamps of each sound event [15].

In this work, we collect audio clips of the AudioGrounding dataset from their original YouTube videos. Because of unavailable YouTube videos, we have 4,253 clips for the training set, 30 clips for the validation set, and 67 clips for the test set. The statistics of our downloaded version for the AudioGrounding dataset are shown in Table 1.

3.2. Experimental Setup

We train the aligning framework with batches of 32 clip-caption pairs in the training set for at most 100 epochs, and monitor the loss (4) on the validation set during the training process. An Adam optimizer with an initial learning rate of 0.001 is adopted to optimize the training process. The learning rate is reduced by a factor of ten once the validation loss does not improve for five epochs. Training is terminated by early stopping with ten epochs.

Split	#Clips	#Captions	#Event phrases
Train	4253	4253	11732
Val	30	150	439
Test	67	335	1118

Table 1. Statistics of the downloaded dataset.

3.3. Evaluation

We evaluate the proposed method with the test set in the AudioGrounding dataset on the tasks of audio-caption retrieval and phrase-based SED.

Audio-Caption Retrieval. This task aims to retrieve instances that are most similar to a given instance from one modality to the other modality, such as retrieving captions pertaining to an audio clip. The task serves to provide a single high-level metric, which captures how well the aligning framework has learned to bridge audio and text modalities at the whole-clip and whole-caption levels. We experiment on audio-to-caption and caption-to-audio retrieval with a clip-wise negative sampling approach, where the correct target caption/clip for a given query clip/caption is always among 29 negative randomly sampled captions/clips. The number of negative samples is determined by the validation size, which only has 30 unique clips. For each query probe in the test set, audio clips and captions are sorted by their clip-caption similarities (3). Retrieval performance is measured in terms of recall at k (R@k with $k \in \{1, 5\}$) averaged across all the audio clips, i.e., checking if the target clip/caption is always within top k search results according to the clip-caption similarity sore. Theoretically, the chance levels are $1/30$ (around 0.03) for R@1 and $1/6$ (around 0.17) for R@5, respectively. To prevent randomness, we repeat the evaluation twenty times

Model		Chance Levels	Proposed
Audio2Caption	R@1	0.03	0.21 ± 0.04
	R@5	0.17	0.65 ± 0.05
Caption2Audio	R@1	0.03	0.23 ± 0.04
	R@5	0.17	0.71 ± 0.04

Table 2. Recall scores of audio-caption retrieval on the test set of the AudioGrounding dataset.

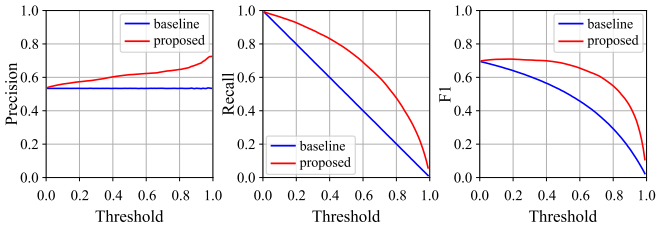

Fig. 3. Frame-based metrics (precision, recall, and F1) against detection threshold for phrase-based SED on the test set of the AudioGrounding dataset.

on the test set.

Phrase-based SED. This task aims to detect sound events within audio clips using sound event phrases in the clips' captions. It serves to explore the ability of the aligning framework on learning correspondences between individual sound events and their textual phrases. Given an audio clip and a sound event phrase from its caption, we experiment with mean and max aggregations to calculate the frame-phrase similarities (5), followed by a min-max normalization. Detection performance is measured with global frame-based metrics (precision, recall, and F1), which are calculated across all clips and sound event phrases. Random guessing is utilized as the baseline that generates frame-phrase similarities with random numbers from a uniform distribution on the interval $[0, 1)$.

3.4. Results and Analysis

Audio-Caption Retrieval. The averages and standard deviations of recall scores with the proposed method are reported in Table 2. In contrast to the theoretical chance levels, the proposed method achieves better recall scores, with R@1 over 0.20 and R@5 over 0.65. The experimental results show that the proposed method can associate audio clips with their corresponding captions, and vice versa.

Phrase-based SED. The frame-based metrics (precision, recall, and F1) against detection threshold are illustrated in Figure 3. The proposed method obtains similar results with either mean or max aggregations in (5), and only the results from mean aggregations are reported. Overall, the proposed method achieves better detection performance than random guessing on the three metrics, regardless of the detection thresholds. Particularly, with a detection threshold of 0.5, the proposed method obtains a precision of 0.62, a recall of 0.77, and a F1 of 0.68 while random guessing has values of 0.53, 0.50, and 0.52, respectively. The experimental results show that the proposed method can detect sound events from audio clips with their textual phrases. Additionally, we experiment with the proposed method by removing min-max normalization from calculating frame-phrase similarities, which results in a drastic reduction in the detection performance.

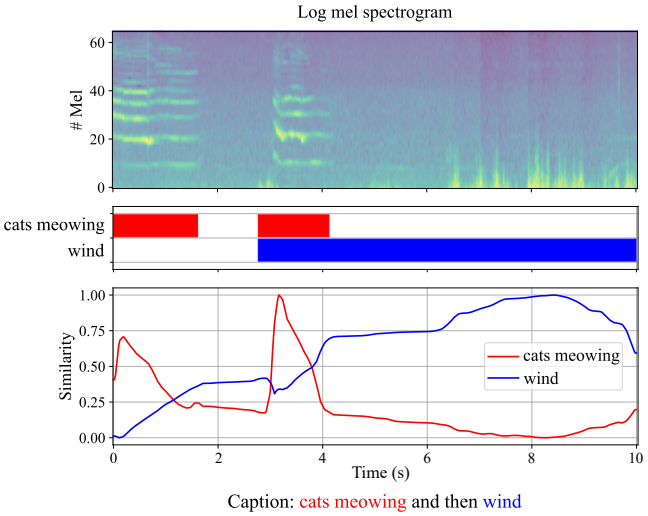

Fig. 4. An example of learned frame-phrase similarities for two sound events and their textual phrases.

An example of learned frame-phrase similarities for two sound events and their corresponding textual phrases is illustrated along time in Figure 4. It shows that high similarities are learned only between sound events and their corresponding textual phrases. We conclude that the proposed method can learn event-phrase correspondences via unsupervised clip-caption aligning.

4. CONCLUSION

We propose an unsupervised method to match sound events with textual phrases by aligning whole captions and audio clips. Audio clips and captions are aligned by scoring the frame-word similarities with their acoustic and semantic embeddings. We evaluate our proposed method with the AudioGrounding dataset [15] on two cross-modal tasks: audio-caption retrieval, and phrase-based SED. Experimental results show that the proposed method can learn global associations between audio clips and captions as well as local correspondences between individual sound events and textual phrases. As future work, we consider exploring large audio-text alignment datasets and fine-tuned semantic embeddings.

5. ACKNOWLEDGEMENT

The research leading to these results has received funding from Emil Aaltonen foundation funded project "Kielen käyttö structuroimattoman datan automaattiseen tulkintaan" and Academy of Finland grant no. 314602.

6. REFERENCES

[1] Benjamin Elizalde, Shuayb Zarar, and Bhiksha Raj, "Cross modal audio search and retrieval with joint embeddings based on text and audio," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2019, pp. 4095–4099.

[2] Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen, "Automated audio captioning with recurrent neural networks," in *Proc. IEEE Workshop Appl. Signal Process. Audio and Acoustic. (WASPAA)*, 2017, pp. 374–378.

[3] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu, "A crnn-gru based reinforcement learning approach to audio captioning," in *Proc. Detect. Classif. Acoust. Scenes Events Work. (DCASE)*, 2019, pp. 225–229.

[4] Haytham M. Fayek and Justin Johnson, "Temporal reasoning via audio question answering," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2283–2294, 2020.

[5] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. (HLT-NAACL)*, 2019, pp. 119–132.

[6] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen, "Clotho: an audio captioning dataset," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2020, pp. 736–740.

[7] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, pp. 423–443, 2019.

[8] David Harwath, Antonio Torralba, and James R. Glass, "Unsupervised learning of spoken language with visual context," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1866–1874.

[9] Zhibin Hu, Yongsheng Luo, Jiong Lin, Yan Yan, and Jian Chen, "Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching," in *Proc. 28th Int. Jt. Conf. Artif. Intell. (IJCAI)*, 2019, pp. 789–795.

[10] Jonatas Wehrmann, Camila Kolling, and Rodrigo C Barros, "Adaptive cross-modal embeddings for image-text alignment," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 12313–12320.

[11] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen, "Cross-modal attention with semantic consistence for image–text matching," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, pp. 5412–5425, 2020.

[12] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass, "Jointly discovering visual objects and spoken words from raw sensory input," *Int. J. Comput. Vis.*, vol. 128, pp. 620–641, 2020.

[13] Khazar Khorrami and Okko Räsänen, "Evaluation of audio-visual alignments in visually grounded speech models," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2021, pp. 2996–3000.

[14] Liming Wang, Xinsheng Wang, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak, "Align or attend? toward more efficient and accurate spoken word discovery using speech-to-image retrieval," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2021, pp. 7603–7607.

[15] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu, "Text-to-audio grounding: Building correspondence between captions and sound events," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2021, pp. 606–610.

[16] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah, "Signature verification using a "siamese" time delay neural network," in *Proc. 6th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 1993, pp. 737–744.

[17] David Harwath, Wei-Ning Hsu, and James R. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *Proc. 8th Int. Conf. Learn. Representations (ICLR)*, 2020, pp. 1–22.

[18] Grzegorz Chrupała, "Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques," pp. 1–29, 2021.

[19] Khazar Khorrami and Okko Räsänen, "Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? — A computational investigation," *Lang. Dev. Res.*, vol. 1, pp. 123–191, 2021.

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Representations (ICLR)*, 2013.

[21] "Word2Vec," https://code.google.com/archive/p/word2vec, Accessed: 2021-09-27.