How Does an Adjective Sound Like? Exploring Audio Phrase Composition with Textual Embeddings

Saba Nazir and **Mehrnoosh Sadrzadeh**
Department of Computer Science
University College London, United Kingdom
saba.nazir.19@ucl.ac.uk, m.sadrzadeh@ucl.ac.uk

Abstract

We learn matrix representations for the frequent sound-relevant adjectives of English and compose them with vector representations of their nouns. The matrices are learnt jointly from audio and textual data, via linear regression and tensor skipgram. They are assessed using an adjective similarity benchmark and also a novel adjective-noun phrase similarity dataset, applied to two tasks: semantic similarity and audio similarity. Joint learning via Tensor Skipgram (TSG) outperforms audio-only models, matrix composition outperforms addition and non compositional phrase vectors.

1 Introduction

Natural language data consists of words arranged into phrases and sentences. Words have statistical representations and phrases/sentences symbolic forms. The formers, mined from co-occurrence counts, fall within the remit of distributional lexical semantics. The latters, often formalised within logic frameworks, are obtained from rules of grammar. A model of natural language should ideally take both into account. Consider a simple adjective-noun phrase. On the lexical side, statistical vector embeddings are learnt for adjectives and nouns. On the symbolic side, e.g. in Combinatory Categorial Grammar (CCG) (Steedman, 2002), an adjective is a function applied to a noun. The lexical and the symbolic sides are brought together by providing a statistical representation for the CCG rules. For the adjective-noun phrase rule, this is achieved by representing adjectives as matrices, nouns as vectors, and function application by matrix-vector multiplication (Baroni and Zamparelli, 2010). This unified model has been applied to multimodal image-text data (Lewis et al., 2022), but never to other combinations such as audio-text. For example, in an audio-text context, adjectives like "loud" or "soft" can modify nouns like "music," where the meaning is enriched by integrating corresponding audio features with their textual representations. Our aim in this paper is to fill this gap. We represent the sounds of adjectives by matrices, the sounds of nouns by vectors, and test whether their matrix-vector multiplication is a good representative of the sound of adjective-noun phrase. To this end, we work with two tasks: a semantic similarity task and an audio similarity one. We develop a new dataset of audio relevant adjective-noun phrases and collect human annotations for them. The matrix representations are from the audio data gathered from FreeSound[1], a collaborative repository of sounds. The correlation between the model's predictions and human annotations is tabulated. These show that matrix-vector adjective-noun composition works better than simple vector addition and non-compositional vectors of adjective-noun phrases. The quality of the audio adjectives significantly improved after auditory and textual data were combined and textual data used as a signal in audio adjective learning. These results show that matrix composition leads to better representations for audio phrases, with potential applications to audio classification (Xie and Virtanen, 2021) and captioning tasks (Mahfuz et al., 2023).

2 Related Work

Using vector addition for composing adjectives with nouns was proposed in (Mitchell and Lapata, 2008). Later, in a series of papers (Grefenstette and Sadrzadeh, 2011; Baroni and Zamparelli, 2010; Maillard and Clark, 2015), it was argued that vector addition is not appropriate for composition as it is commutative. Furthermore, an adjective needs to *modify* the meaning of a noun, thus its representation should be a map, rather than a vector. In finite dimensions, maps are approximated by matrices and adjective-noun phrase composi-

[1] https://freesound.org

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 13–18
October 14–15, 2024. ©2024 Association for Computational Linguistics

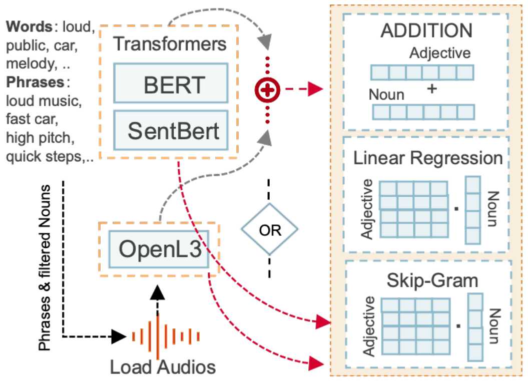

Figure 1: For audio vectors, we used the pre-trained OpenL3 (Cramer et al., 2019) library, trained on environmental and musical data from AudioSet (Gemmeke et al., 2017). OpenL3 uses a convolutional architecture initialised on a Mel-spectrogram time-frequency representation with 256 bands; its vectors are 512 dimensional. For textual vectors, we used 768 dimensional pre-trained BERT embeddings (Devlin et al., 2018) for words and SBERT (Reimers and Gurevych, 2019) for phrases.

tion becomes matrix-vector multiplication, a non-commutative operation. Different methodos were put forwards for learning the adjective matrices; (Baroni and Zamparelli, 2010) used linear regression and (Maillard and Clark, 2015; Wijnholds and Sadrzadeh, 2019) developed a tensorial extension of the word2vec skipgram model (Mikolov et al., 2013). Learning multimodal image-text embeddings for words was proposed in (Bruni et al., 2014; Lazaridou et al., 2015); extended to sound-text in (Kiela and Clark, 2015). Matrix composition of images and text was explored in (Lewis et al., 2022).

3 Single and Multi Modal Learning

An overview of multimodal phrase composition is presented in Figure 1. To learn the matrices, we used linear regression (LR) and the tensorial extension of skipgram (TSG). For LR, we trained adjective matrices $\boldsymbol{A}$ given observed adjective-noun vectors $\boldsymbol{p}$ and noun vectors $\boldsymbol{v}$, using the formula $\boldsymbol{p} = \boldsymbol{A}\boldsymbol{v}$.

The original word2vec skipgram model had the following objective function, where $\boldsymbol{n}$ is a vector, and $\mathcal{C}$ and $\overline{\mathcal{C}}$ sets of positive and negative contexts.

$$\sum_{\boldsymbol{c}'\in\mathcal{C}} \log\sigma\left(\boldsymbol{w}\cdot\boldsymbol{c}'\right) + \sum_{\overline{\boldsymbol{c}}'\in\overline{\mathcal{C}}} \log\sigma\left(-\boldsymbol{w}\cdot\overline{\boldsymbol{c}}'\right)$$

This model learns a vector for a word w regardless of its grammatical type. Its tensorial extension, dubbed as **tensor skipgram** has an objective function that depends on the grammatical role of the words. For adjective-noun phrases, this is as follows, where $\mathbf{A}$ is the adjective matrix, $\boldsymbol{n}$ the vector of the noun it modifies, and the rest is as before.

$$\sum_{\boldsymbol{c}'\in\mathcal{C}} \log\sigma\left(\mathbf{A}\boldsymbol{n}\cdot\boldsymbol{c}'\right) + \sum_{\overline{\boldsymbol{c}}'\in\overline{\mathcal{C}}} \log\sigma\left(-\mathbf{A}\boldsymbol{n}\cdot\overline{\boldsymbol{c}}'\right)$$

The above function is only for adjective-noun phrases. It generalises to any phrase in (Wijnholds and Sadrzadeh, 2019). TSG significantly outperforms LR on text (Maillard and Clark, 2015; Wijnholds and Sadrzadeh, 2019).

The audio and textual representations were combined with two different methods. In the first method, we concatenated their vectors (**AT-Concat**) and used the result as an input to training. In the second method, we trained a joint audio-text matrix (**AT-Joint**), where one representation was used as a signal to improve the other.

AT-Concat Regression uses the following adaptation of the above single modality regression:

$$\langle \boldsymbol{p}^a, \boldsymbol{p}^t \rangle = \mathbf{A}\langle \boldsymbol{v}^a, \boldsymbol{v}^t \rangle$$

where $\boldsymbol{v}^a$ is the audio representation of a noun, $\boldsymbol{v}^t$ its textual counterpart, and $\langle \boldsymbol{v}^a, \boldsymbol{v}^t \rangle$ their concatenation. Similarly, $\boldsymbol{p}^a$ is the audio representation of an adjective-noun phrase, $\boldsymbol{p}^t$ its textual counterpart, and $\langle \boldsymbol{p}^a, \boldsymbol{p}^t \rangle$ their concatenation.

AT-Joint Regression uses the following variant of the original regression formula $\boldsymbol{p}^a = \mathbf{A}\boldsymbol{v}^t$ for training, where the audio adjective-noun phrase vector $\boldsymbol{p}^a$ uses the textual representation of its noun $\boldsymbol{v}^t$ as a signal to learn an adjective matrix $\mathbf{A}$, which has a combined audio-text meaning.

AT-Concat Tensor Skipgram is based on the modified training objective of the single modality TSG and has the following objective function (to save space we only provide the positive sampling part):

$$\sum_{(\boldsymbol{c'^a}, \boldsymbol{c'^t}) \,\in\, \mathcal{C}^a \times \mathcal{C}^t} \log\sigma\left(\mathbf{A}\langle \boldsymbol{n^a}, \boldsymbol{n^t} \rangle \cdot \langle \boldsymbol{c'^a}, \boldsymbol{c'^t} \rangle\right)$$

Here, $\langle \boldsymbol{n}^a, \boldsymbol{n}^t \rangle$ is the concatenation of the fixed pretrained audio and textual embeddings of a noun, and $\mathcal{C}^a, \mathcal{C}^t$ are sets of positive and negative contexts of the adjective-noun phrase. For positive contexts, we use the fixed pretrained embeddings of the actual audio and text representations of the adjective-noun phrases. For negative contexts, we fix the adjective and randomly chose a subset of nouns different from n. For example, to learn a matrix $\mathbf{A}$ for the adjective *happy*, $\boldsymbol{n}^t$ is the textual embedding of *cat* and $\boldsymbol{n}^a$ the average of all its audio vectors; $\boldsymbol{c'^a}$ indexes over all the audio embeddings we have for *happy cat* and $\boldsymbol{c'^t}$ is its textual embedding. For negative contexts, $\overline{\boldsymbol{c}}'^{\boldsymbol{a}}$ indexes over all the audio embeddings we have for *happy noun*, where *noun* is a random noun different from *cat*, e.g. *baby* and *car*.

AT-Joint Tensor Skipgram changes the objective function to the following, for the same $\boldsymbol{n}^t$ and $\mathcal{C}^a$ as above.

$$\sum_{\boldsymbol{c'^a} \in \mathcal{C}^a} \log \sigma \left(\mathbf{A}\boldsymbol{n}^t \cdot \boldsymbol{c'^a} \right) + \sum_{\overline{\boldsymbol{c}}'^a \in \overline{\mathcal{C}}} \log \sigma \left(-\mathbf{A}\boldsymbol{n}^t \cdot \overline{\boldsymbol{c}}'^a \right)$$

Here, the audio adjective is learnt from an audio-only context, but in such a way that when multiplied with the textual vector of a noun, it is forced to be closer to the audio context.

4 Implementation

We implemented an audio-text TSG, by extending the image-text TSG model of (Lewis et al., 2022) to audio data. The positive context is the audio files representing a target phrase. For instance, for *loud melody* we had 100 audio files and for *loud cat* 82. The negative context is determined by random selection of nouns during the training process with each adjective. We treat these nouns as a hyper parameter and choose them by tuning on the validation segment of the dataset.

For skipgram models, we learn 50 dimensional phrase vectors with a learning rate of 10^{-6} and a batch size of 512, trained for 200 epochs. The models were trained on NVIDIA T4 and V100 depending on their availability on Google Colab. The training was done in batches over a period of 3 months, totalling 80 hrs. We used Binary Cross-Entropy loss and the Adam optimiser in the training process to refine the performance.

5 Evaluation Tasks and Results

Our main hypothesis is that combining text and audio improves over audio-only learning. To test this, we trained audio-only variants of LR and TSG models. In these, the adjective matrices were learnt using only the audio vectors of their nouns and contexts. A second hypothesis is that non-commutative matrix multiplication models (LR and TSG) outperform simple commutative models. To test this, we implemented an additive model where an adjective's representation is added to its nouns. Finally, we hypothesise that compositional models outperform non-compositional ones. For this, we compared the results to the holistic OpenL3 audio vector of adjective-noun phrases.

5.1 Adjective Similarity

Following (Maillard and Clark, 2015), we first evaluate our methods on an adjective similarity task. Starting from the word similarity dataset SimLex-999 (Hill et al., 2015), We identified 13 sound-relevant (adj, adj) pairs with audio files in FreeSound. These pairs represent 11 out of 30 adjectives from our dataset. We call it *Simlex-Audio*. Examples are (*happy, cheerful*) and (*fast, rapid*).

5.2 Adjective-Noun Similarity

Existing adjective-noun phrase similarity benchmarks, such as (Mitchell and Lapata, 2010; Vecchi et al., 2017) were unsuitable due to limited sound relevance. This led us to develop a new audio phrase dataset.We selected frequent *audio adjectives* from the UKWaC corpus (top 1000 adjectives with at least 200 occurrences) and those with strong auditory relevance in FreeSound (800+ mentions)[2], resulting in 30 suitable adjectives, each paired with a noun. Nouns were refined grammatically and filtered to those with 100+ mentions on Freesound.

This procedure resulted in a dataset of 30 adjectives, 721 unique nouns, and 1,944 adjective-noun phrases. The number of nouns modified by each adjective varied; for example, *low* modified 46 nouns, while *quick* modified 114, with an average of 65 nouns per adjective. For audios, we selected 100 audio files per noun and on average 50 files per adjective-noun phrase, each 10-20 seconds long. The number of audio files per adjective-noun varied, e.g., *human cough* had 97 audios and *angry girl* had 45. The dataset contained 271,766 files (about 760 hours), split into 80% training, 10% testing, and 10% validation for experimentation.

5.3 Semantic and Audio Similarity Tasks

The new audio phrase dataset includes both semantic and audio similarity judgments, scored from 1 (least similar) to 5 (most similar). Annotators scored pairs based on semantic relatedness and perceived sound similarity. A pilot study with 100 randomly chosen phrase pairs and 10 annotators yielded an inter-annotator agreement of 0.45. To improve this, pairs with identical adjectives were categorized as *environmental* (e.g., *happy cat*, *loud wind*) or *musical* (e.g., *loud piano*). The data was

[2] We refer to these adjectives as audio-relevant due to their strong association with sounds.

Model	**Adjective Similarities**		**Phrase Similarities**			
	Simlex-Audio		**SemPhrase**		**AudPhrase**	
	LR	**TSG**	**LR**	**TSG**	**LR**	**TSG**
AT-Concat	0.731	0.755	0.762	0.856	0.779	0.876
AT-Joint	0.635	0.79	0.668	0.882	0.581	0.894
Audio-Only	0.683	0.743	0.716	0.783	0.753	0.825
ADD-Audio	0.455		0.689		0.743	
ADD-AT	0.499		0.647		0.669	
Non-Comp Audio	–		0.511		0.578	

Table 1: Similarities computed for Simlex-Audio, SemPhrase, and AudPhrase datasets. Non-Comp, ADD, LR, and TSG denote Non-Compositional, Addition, Linear Regression, and Tensor Skipgram; **AT** is Audio-Text, and **Concat** is concatenation.

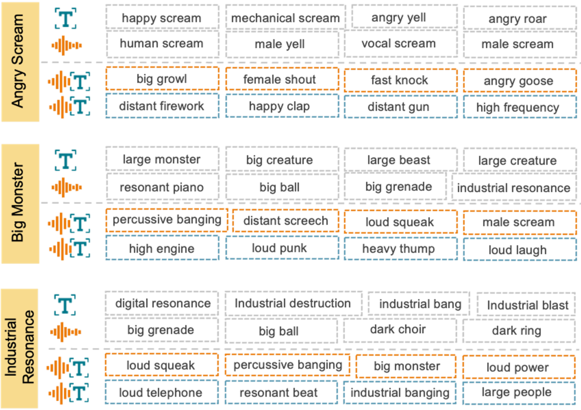

Figure 2: Query and its top 4 closely related phrases(left to right). Grey rows indicate non-comp audio and text-based similarities, while orange and blue signify similar phrases for compositional audio and semantic similarities, using AT-Joint.

arranged into forms of 10 pairs; each with only either musical or environmental phrases. 4 forms were grouped together to create 1 questionnaire.

Human Judgements: We used Amazon Mechanical Turk to collect annotations, selecting annotators with a HIT approval rate above 95% and over 1000 approved HITs. They were paid £10.42/hr. Tasks were batched with gold standards to filter automated responses, excluding unexpectedly fast annotations. To manage costs, we limited the nouns per adjective to 15-20, with 100 sound files each. This resulted in 3,144 adjective-noun pairs across 77 questionnaires, each annotated by 15 different annotators, totalling 113. Inter-annotator agreement was 0.69 for semantic similarity and 0.67 for audio similarity. We call these datasets SemPhrase and AudPhrase and they will be available on github[3].

5.4 Results

We measured the Spearman correlation ρ_s between the human annotations and cosine similarities, see Table 1 for the results. For semantic similarity and in both SimLex and our new dataset, the best performing model was the audio-text joint learning (**AT-Joint**) via TSG. The second best performing model was audio-text concatenation (**AT-Concat**) via TSG. They both improved on their LR counterparts, and outperformed the audio-only, additive, and non compositional models. In LR, only **AT-Concat** outperformed all the baselines; but itself fell short of TSG. A very similar trend was observed for the audio similarity task, where TSG applied to **AT-Joint** was the best performing model again, outperforming all baselines. The second best model was TSG with **AT-Concat**. For LR, again only **AT-Concat** outperformed the baselines.

6 Discussion and Conclusion

We conducted a case study to understand the better performance of compositional multimodal audio-text embeddings using k-means clustering with optimal k values determined via Silhouette method (Rousseeuw, 1987). Cosine similarities were computed within each cluster, to find closet neighbours to the random queries from the evaluation split. Some examples are provided in Figure 2. We found out that holistic singular text and audio only models predicted either semantic or audio relevance, often getting close to opposite concepts or literal sounds. On the other hand, multimodal composition managed to predict a more accurate phrase meaning. When non-compositional models struggle to predict, e.g. in the second example, the audio-only model predicted *resonant piano* and *big ball* as synonyms of *big monster*, multimodal composition predicted *loud squeak* and *heavy thump* and bridged the gap. Another example is the prediction of *distant firework*, *distant gun*, and *high frequency* for *angry scream* by multimodal composition, where a text-only model guessed the opposite, i.e. *happy scream*.

Similar is the case for *industrial resonance*, predicted to be close to *percussive banging* and *loud telephone* by the compositional model, improving over the audio-only model which predicted *big monster* and the text-only model which again predicted opposite, i.e. *industrial blast*.

These findings show that reflecting the textual grammatical structure in adjective-noun composition and considering both audio and text modalities improves the quality of audio data. Extending the setting to verb phrases is work in progress.

[3] https://github.com/audio-comp

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1183–1193.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. 2019. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Douwe Kiela and Stephen Clark. 2015. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2461–2470.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Martha Lewis, Qinan Yu, Jack Merullo, and Ellie Pavlick. 2022. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*.

Rehana Mahfuz, Yinyi Guo, and Erik Visser. 2023. Improving audio captioning using semantic similarity metrics. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jean Maillard and Stephen Clark. 2015. Learning adjective meanings with a tensor-based skip-gram model. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 327–331.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Mark Steedman. 2002. Mark steedman, the syntactic process (language, speech, and communication). cambridge, ma: Mit press, 2000. pp. xiv 330. *Journal of Linguistics*, 38(3):645–708.

Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek. 2020. Pre-training audio representations with self-supervision. *IEEE Signal Processing Letters*, 27:600–604.

Eva M Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2017. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive science*, 41(1):102–136.

Gijs Wijnholds and Mehrnoosh Sadrzadeh. 2019. Evaluating composition models for verb phrase elliptical sentence embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 261–271, Minneapolis, Minnesota. Association for Computational Linguistics.

Huang Xie and Tuomas Virtanen. 2021. Zero-shot audio classification via semantic embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1233–1242.