How Does an Adjective Sound Like? Improving Audio Phrase Composition with Text Embeddings

Anonymous ACL submission

Abstract

We learn matrix representations for the most 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 (LR) and tensor skipgram (TSG). Their quality is as assessed on a novel adjective-noun phrase similarity dataset, applied to two tasks: semantic similarity and audio similarity. Joint learning via 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 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. 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, a collaborative repository of sounds[1]. The correlation between the predictions of the model and the annotations provided by humans are 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 composition becomes matrix-vector multiplication, a non-commutative operation. Different methodologies 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 exten-

[1] https://freesound.org

sion 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); it was 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

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.

To learn the matrices, we used linear regression and the tensorial extension of skipgram. For linear regression, 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). Tensor skipgram significantly outperforms regression 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 Tensor skipgram 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 tensor skipgram, by extending the image-text tensor skipgram model of (Lewis et al., 2022) to audio data. The positive context is fixed and is defined as the number of audio files representing a target phrase. For instance, for *loud melody* we had 100 and for *loud cat* 82. Conversely, the negative context is determined by random selection of nouns during the training process within 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, the learning rate was 10^{-6} with a batch size of 512, and a training duration of 200 epochs. The models were trained on NVIDIA T4 and V100 depending on their availability on Google Collab. The training was done in batches over a period of 3 months, totalling 100 hrs. We used Binary Cross-Entropy loss and the Adam optimiser in the training process to refine the performance. Principal Component Analysis (PCA) was used to equalize the dimensions of auditory and textual representations to 50.

5 Dataset

Traditional adjective-noun phrase similarity benchmarks, such as (Mitchell and Lapata, 2010; Vecchi et al., 2017) were unsuitable for our study due to their limited sound relevance: most of the adjectives and nouns of these datasets did not have any sound files in FreeSound. Further, their entries had different adjectives. Therefore, we had to form a new own dataset by first choosing a set of audio-relevant adjectives, then forming adjective-noun phrases from them.

The chosen adjectives had both a high frequency of usage in English and a strong relevance to auditory experiences. For English usage, we used the UKWaC (Ferraresi et al., 2008) corpus, and for auditory experience, the Freesound library. We refer to the resulting adjectives as *audio adjectives*. They were collected as follows: first we found UKWaC's 1000 most frequent adjectives that had occurred no fewer than 200 times. Next, we searched the Freesound file names and tags for these adjectives and kept those that had 800 or more instances. We only chose adjectives that were accompanied by a noun. This resulted in 30 adjectives.

The nouns that these adjectives had modified were retrieved after a post processing step. This had two substeps: (a) a textual step, which involved singularizing plurals, correcting nouns via a spellchecker, and manually removing ambiguous and nonsensical nouns such as *file*, (b) an auditory step, where the Freesounds library was searched with the nouns and only those that had at least 100 occurrences in file names or tags were kept.

This procedure resulted in a dataset of 30 adjectives, 721 unique nouns, and 1,944 adjective-noun phrases. The number of nouns each adjective modified varied; for instance, for the adjective *low* we found 46 sound-relevant nouns, but for adjective *quick* 114. On average, each adjective modified approximately 65 nouns. Each noun and adjective-noun had many corresponding audio files. The number of files per noun varied from 100's to 1000's; from these we randomly chose the 100 that had a length between 10 to 20 seconds. For each adjective-noun, the number of files varied from 10 to 100, from which we again randomly chose 50 files of length 10-20 secs. For instance, the phrase *human cough* had 97 sound files, whereas for *angry girl* we only found 45. In total, our dataset had 271,766 audio files, equivalent to approximately 760 hours of audio. We allocated 80% of the dataset to training, 10% to test, and 10% to validation.

6 Evaluation Tasks and Results

Our main hypothesis is that concatenation and joint learning of text and audio should improve on audio-only learning. In order to test this hypothesis, we also trained audio-only variants for both regression and tensor skipgram models. In these, the adjective matrices were learnt by using only the audio vectors of their nouns and contexts. A second hypothesis is that non-commutative matrix multiplication models, i.e. regression and tensor skipgram, should improve on simple commutative models. In order to test this hypothesis, we implement an additive model, where the representation of an adjective was added to that of its noun. Our final hypothesis is that compositional models should outperform non-compositional ones. For this, we compared the results to the holistic OpenL3 audio vector of adjective-noun phrases.

6.1 Semantic and Audio Similarity Tasks

We collected two different types of human judgements: one for a semantic similarity task and another for an audio similarity task. Each judgement was a score between 1 and 5, where 1 stood for least

Model	Semantic Similarity	
	LR	TSG
AT-Concat	0.762	0.856
AT-Joint	0.668	0.882
Audio-Only	0.716	0.783
ADD-Audio	0.689	
ADD-AT	0.647	
Non-Comp Audio	0.511	

Model	Audio Similarity	
	LR	TSG
AT-Concat	0.779	0.876
AT-Joint	0.581	0.894
Audio-Only	0.753	0.825
ADD-Audio	0.743	
ADD-AT	0.669	
Non-Comp Audio	0.578	

Table 1: Tables of Results. Non-Comp, ADD, LR and TSG represent Non-Compositional, Addition, Linear Regression and tensor skipgram; **AT** is for Aduio-Text, **Con** for concatenation, **Joint** for joint learning.

similar and 5 for most similar. In the semantic similarity task, we asked the annotators to score each pair based on the semantic relatedness of its entries. In the audio similarity, we asked for a score on how similar the sounds of the entries were. A pilot study with 100 pairs of randomly chosen phrases and 10 annotators resulted in an inter-annotator agreement of 0.45. In order to improve on this, the pairs were restricted to those with identical adjectives and categorised into *environmental* or *musical*. An example of a musical phrase was *loud piano*, examples of environmental phrases were *happy cat* and *loud wind*. The data was arranged into forms of 10 pairs; each with only either musical or environmental phrases. 4 forms were grouped together to create 1 questionnaire.

6.2 Human Judgements

We launched the tasks on Amazon's Mechanical Turk platform and collected annotations from English-speaking countries with a HIT approval rate greater than 95% and the number of approved HITs greater than 1000. The annotators were paid £10.42 per hour (the minimum UK wage). The data of each task was divided into batches and each batch had a few gold standards to help identify automated responses. Additionally, the time used per task by each annotator was recorded, and if an annotator completed a task significantly faster than expected, their annotation was excluded. In order to keep the expenses at a reasonable level, the number of nouns per adjective was restricted to 15-20, which were the ones that had exactly 100 sound files. This resulted in 3,144 pairs of adjective-noun phrases, grouped into 77 questionnaires. Each questionnaire was annotated by 15 different annotators totalling 113. Inter-annotator agreement was 0.69 for semantic similarity and 0.67 for audio similarity. The annotations would shortly be available on GitHub[2].

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

6.3 Results

We measured the Spearman correlation ρ_s between the human annotations and cosine similarities, see Table 1 for the results. For semantic similarity, the best performing model was the audio-text joint learning (**AT-Joint**) via tensor skipgram (TSG). The second best performing model was audio-text concatenation (**AT-Concat**) via TSG. They both improved on their linear regression (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 held for the audio similarity task, where TSG applied to **AT-Joint** was the best performing model again, outperforming all baselines. The second best performing model was TSG applied to **AT-Concat**. For LR, again only **AT-Concat** outperformed the baselines.

We performed a qualitative analysis for our best performing model **AT-Joint** TSG. Here, we looked the nearest neighbours of a sample of randomly chosen phrases. Here are some examples. In semantic similarity, *angry boss* was closest to *angry person*, *deep punch*, and *big hammer*, forming a group related to anger and terror. In audio similarity, it was closest to *distant screech*, *big groan*, and *big noise*, capturing the sound-related aspects of the concept of anger.

7 Conclusion

We learnt matrices for the audio data of adjectives and composed them with the audio embeddings of their nouns. The quality of the audio adjectives improved in a multimodal setting and when textual data was injected into the learning procedure. This shows the grammatical structure reflected in the adjective-noun phrase composition in text also holds for audio data. Extending the setting to verb phrases and whole sentences is work in progress.

8 Limitations

Adjective Similarity Tasks We evaluated our methods on phrase similarity tasks. It, however, does make sense to also evaluate them on adjective similarity tasks. The original single modality variants of these methods have been applied to adjective similarity tasks, see (Maillard and Clark, 2015). The limitation we faced was that there was a very small overlap between the audio-relevant adjective-only subsets of the existing word similarity datasets. The largest overlap was with SimLex (Hill et al., 2015), which had 10 audio-relevant adjective-only pairs. We evaluated our methods on these few pairs. For both audio and semantic similarity, the audio-text model outperformed the audio-only model, with the difference that here (as opposed to the adjective-noun similarity tasks) **AT-Concat** (and not TSG) was the best. In order to overcome this limitation, one needs to develop a new dataset of audio-relevant adjective pairs and collect human judgements for it. The challenge is to find the adjectives for which semantic or audio similarity would make sense.

Using audio as signal to text Our focus on this paper was to make sense of adjective-noun composition in audio data, and how a text signal can improve on this. It would also make sense to explore a different variant of this question: whether text representations of adjectives can be improved by using audio data. In order to address this question, we need to learn the adjective matrices with a different set of objectives, i.e. the one below for regression

$$p^t = \mathbf{A} \times v^a$$

and the one below from tensor skipgram:

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

The immediate challenge faced when attempting to implement the above was lack of enough data. A text context $\boldsymbol{c'^t}$ is only one vector, where as an audio context, e..g the one previously used $\boldsymbol{c'^a}$ consists of many (in this paper up to 100) sound files. This challenge can be overcome by considering many text contexts, for instance by working with semantically similar nouns to $\boldsymbol{c'^t}$ or using temporal recurrent neural networks (Tagliasacchi et al., 2020). These directions constitute work in progress.

Text-Only Concatenation and Text-Only Joint Learning It is possible to compare our results with a text-only model where the LR and TSG methods are implemented on text-only corpora such as the UKWaC. We are at the moment training and fine-tuning these models and hope to be able to present the results in another paper.

Middle and late Concatenation We did not implement separate textual models and primarily engaged in joint learning of audio-textual models. As a result, our methodologies align with early fusion of modalities, Multimodal text-image and text-audio learning have also been developed for middle and late fusion approaches. These can be investigated when we implement the methods on textual only.

Other Applications We only dealt with adjective-noun similarity. The model can be extended to full sentences where it can be applied to a range of other applications such audio captioning.

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.

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.