Multi- and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception

Douwe Kiela
Computer Laboratory
University of Cambridge
douwe.kiela@cl.cam.ac.uk

Stephen Clark
Computer Laboratory
University of Cambridge
stephen.clark@cl.cam.ac.uk

Abstract

Multi-modal semantics has relied on feature norms or raw image data for perceptual input. In this paper we examine grounding semantic representations in raw auditory data, using standard evaluations for multi-modal semantics, including measuring conceptual similarity and relatedness. We also evaluate cross-modal mappings, through a zero-shot learning task mapping between linguistic and auditory modalities. In addition, we evaluate multi-modal representations on an unsupervised musical instrument clustering task. To our knowledge, this is the first work to combine linguistic and auditory information into multi-modal representations.

1 Introduction

Although distributional models (Turney and Pantel, 2010; Clark, 2015) have proved useful for a variety of NLP tasks, the fact that the meaning of a word is represented as a distribution over other words implies that they suffer from the *grounding problem* (Harnad, 1990); i.e. they do not account for the fact that human semantic knowledge is grounded in the perceptual system (Louwerse, 2008). Motivated by human concept acquisition, multi-modal semantics enhances linguistic representations with extra-linguistic perceptual input. These models outperform language-only models on a range of tasks, including modelling semantic similarity and relatedness, and predicting compositionality (Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013; Bruni et al., 2014). Although feature norms have also been used, raw image data has become the de-facto perceptual modality in multi-modal models.

However, if the objective is to ground semantic representations in perceptual information, why stop at image data? The meaning of *violin* is surely not only grounded in its visual properties, such as shape, color and texture, but also in its sound, pitch and timbre. To understand how perceptual input leads to conceptual representation, we should use as many perceptual modalities as possible. A recent preliminary study by Lopopolo and van Miltenburg (2015) found that it is possible to derive uni-modal semantic representations from sound data. Here, we explore taking multi-modal semantics beyond its current reliance on image data and experiment with grounding semantic representations in the auditory perceptual modality.

Multi-modal models that rely on raw image data have typically used "bag of visual words" (BoVW) representations (Sivic and Zisserman, 2003). We follow a similar approach for the auditory modality and construct bag of audio words (BoAW) representations. Following previous work in multi-modal semantics, we evaluate these models on measuring conceptual similarity and relatedness, and inducing cross-modal mappings between modalities to perform zero-shot learning. In addition, we evaluate on an unsupervised musical instrument clustering task. Our findings indicate that multi-modal representations enriched with auditory information perform well on relatedness and similarity tasks, particularly on words that have auditory assocations. To our knowledge, this is the first work to combine linguistic and auditory representations in multi-modal semantics.

2 Related Work

Information processing in the brain can be roughly described to occur on three levels: perceptual input, conceptual representation and symbolic reasoning (Gazzaniga, 1995). While research in AI has made great progress in understanding the first and last of these, understanding the middle level is still more of an open problem: how is it that per-

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2461–2470,
Lisbon, Portugal, 17-21 September 2015.

ceptual input leads to conceptual representations that can be processed and reasoned with?

A key observation is that concepts are, through perception, *grounded* in physical reality and sensorimotor experience (Harnad, 1990; Louwerse, 2008), and there has been a surge of recent work on perceptually grounded semantic models that try to account for this fact. These models learn semantic representations from both textual and perceptual input, using either feature norms (Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013; Hill and Korhonen, 2014) or raw image data (Feng and Lapata, 2010; Leong and Mihalcea, 2011; Bruni et al., 2014) as the source of perceptual information. A popular approach in the latter case is to collect images associated with a concept, and then lay out each image as a set of keypoints on a dense grid, where each keypoint is represented by a robust local feature descriptor such as SIFT (Lowe, 2004). These local descriptors are subsequently clustered into a set of "visual words" using a standard clustering algorithm such as k-means and then quantized into vector representations by comparing the descriptors with the centroids. An alternative to this bag of visual words (BoVW) approach is transferring features from convolutional neural networks (Kiela and Bottou, 2014).

Various ways of aggregating images into visual representations have been proposed, such as taking the mean or the elementwise maximum. Ideally, one would jointly learn multi-modal representations from parallel multi-modal data, such as text containing images (Silberer and Lapata, 2014) or images described with speech (Synnaeve et al., 2014), but such data is hard to obtain, has limited coverage and can be noisy. Hence, image representations are often learned independently. Aggregated visual representations are subsequently combined with a traditional linguistic space to form a multi-modal model. This mixing can be done in a variety of ways, ranging from simple concatenation to more sophisticated fusion methods (Bruni et al., 2014).

Cross-modal semantics, instead of being concerned with improving semantic representations through grounding, focuses on the problem of reference. Using, for instance, mappings between visual and textual space, the objective is to learn which words refer to which objects (Lazaridou et al., 2014). This problem is very much related to the object recognition task in computer vision, but instead of using just visual data and labels, these cross-modal models also utilize textual information (Socher et al., 2014; Frome et al., 2013). This allows for *zero-shot learning*, where the model can predict how an object relates to other concepts just from seeing an image of the object, but without ever having previously encountered an image of that particular object (Lazaridou et al., 2014). Multi-modal and cross-modal approaches have outperformed state-of-the-art text-based methods on a variety of tasks (Bruni et al., 2014; Silberer and Lapata, 2014).

MEN	score	SimLex-999	score
automobile-car	1.00	taxi-cab	0.92
rain-storm	0.98	plane-jet	0.81
cat-feline	0.96	horse-mare	0.83
jazz-musician	0.88	sheep-lamb	0.84
bird-eagle	0.88	bird-hawk	0.79
highway-traffic	0.88	band-orchestra	0.71
guitar-piano	0.86	music-melody	0.70

Table 1: Examples of pairs in the datasets where auditory is relevant, with the similarity score.

3 Evaluations

Following previous work in multi-modal semantics, we evaluate on two standard similarity and relatedness datasets: SimLex-999 (Hill et al., 2014) and the MEN test collection (Bruni et al., 2014). These datasets consist of concept pairs together with a human-annotated similarity or relatedness score, where the former dataset focuses on genuine similarity (e.g., *teacher-instructor*) and the latter focuses more on relatedness (e.g., *river-water*). In addition, following previous work in cross-modal semantics, we evaluate on the zero-shot learning task of inducing a cross-modal mapping to the correct label in the auditory modality from the linguistic one and vice-versa.

3.1 Multi-modal Semantics

Evidence suggests that the inclusion of visual representations only improves performance for certain concepts, and that in some cases the introduction of visual information is detrimental to performance on similarity and relatedness tasks (Kiela et al., 2014). The same is likely to be true for other perceptual modalities: in the case of comparisons such as *guitar-piano*, the auditory modal-

Dataset	MEN	AMEN	SLex	ASLex
Linguistic	3000	258	999	296
Auditory	2590	233	534	216

Table 2: Number of concept pairs for which representations are available in each modality.

ity is certainly meaningful, whereas in the case of *democracy-anarchism* it is probably less so. Therefore, we had two graduate students annotate the datasets according to whether auditory perception is relevant to the pairwise comparison. The annotation criterion was as follows: if both concepts in a pairwise comparison have a distinctive associated sound, the modality is deemed relevant. Inter-annotator agreement was high: $\kappa = 0.93$ for MEN and $\kappa = 0.92$ for SimLex-999. Some examples of relevant pairs can be found in Table 1. Hence, we now have four evaluation datasets: the MEN test collection **MEN** and its auditory-relevant subset **AMEN**; and the SimLex-999 dataset **SLex** and its auditory-relevant subset **ASLex**. Due to the nature of the auditory data sources, it is not possible to build auditory representations for all concepts in the test sets. Hence, unless stated otherwise, we report results for the covered subsets (using the same subsets when comparing across modalities, to ensure a fair comparison). Table 2 shows how much of the test sets are covered for each modality.[1]

3.2 Cross-modal Semantics

In addition to evaluating our models on the MEN and SimLex tasks, we evaluate on the cross-modal task of zero-shot learning. In the case of vision, Lazaridou et al. (2014) studied the possibility of predicting from "we found a cute, hairy wampimuk sleeping behind the tree" that a "wampimuk" will probably look like a small furry animal, even though a wampimuk has never been seen before. We evaluate zero-shot learning, using partial least squares regression (PLSR) to obtain cross-modal mappings from the linguistic to auditory space and vice versa.[2] Thus, given a linguistic representation for e.g. *guitar*, the task is to map it to the appropriate place in auditory space without ever having heard a guitar; or map it to the appropriate place in linguistic space without ever having read about a guitar (having only heard it).

4 Approach

One reason for using raw image data in multi-modal models is that there is a wide variety of resources that contain tagged images, such as ImageNet (Deng et al., 2009) and the ESP Game dataset (Von Ahn and Dabbish, 2004). However, such resources do not exist for audio files, and so we follow a similar approach to Fergus et al. (2005) and Bergsma and Goebel (2011), who use Google Images to obtain images. We use the online search engine Freesound[3] to obtain audio files. Freesound is a collaborative database released under Creative Commons licenses, in the form of snippets, samples and recordings, that is aimed at sound artists. The Freesound API allows users to easily search for audio files that have been tagged using certain keywords.

For each of the concepts in the evaluation datasets, we used the Freesound API to obtain samples encoded in the standard open source OGG format[4]. Because the database contains variable numbers of files, with varying duration per individual file, we restrict the search to a maximum of 50 files and a maximum of 1 minute duration per file. The Freesound API allows for various degrees of keyword matching: we opted for the strictest keyword matching, in that the audio file needs to have been purposely tagged with the given word (the alternative includes searching the text description for matching keywords). For example, if we are searching for audio files of cars, we retrieve up to 50 files with a maximum duration of 1 minute per file that have been tagged with the label "car".

4.1 Linguistic Representations

For the linguistic representations we use the continuous vector representations from the log-linear skip-gram model of Mikolov et al. (2013). Specifically, we trained 300-dimensional vector representations trained on a dump of the English Wikipedia plus newswire (8 billion words in total).[5] These types of representations have been found to yield the highest performance on a variety of semantic similarity tasks (Baroni et al., 2014).

[1] To facilitate further work in multi-modal semantics beyond vision, our code and data have been made publicly available at http://www.cl.cam.ac.uk/~dk427/audio.html.

[2] To avoid introducing another parameter, we set the number of latent variables in the cross-modal PLSR map to a third of the number of dimensions of the perceptual representation.

[3] http://www.freesound.org.

[4] http://www.vorbis.com.

[5] We used the demo-train-big-model-v1.sh script from http://word2vec.googlecode.com to obtain this corpus.

4.2 Auditory Representations

A common approach to obtaining acoustic features of audio files is the Mel-scale Frequency Cepstral Coefficient (MFCC) (O'Shaughnessy, 1987). MFCC features are abundant in a wide variety of applications in audio signal processing, ranging from audio information retrieval, to speech and speaker recognition, and music analysis (Eronen, 2003). Such features are derived from the mel-frequency cepstrum representation of an audio fragment (Stevens et al., 1937). In MFCC, frequency bands are spaced along the mel scale, which has the advantage that it approximates human auditory perception more closely than e.g. linearly-spaced frequency bands. Hence, MFCC takes human perceptual sensitivity to audio frequencies into consideration, which makes it suitable for e.g. compression and recognition tasks, but also for our current objective of modelling auditory perception. We obtain MFCC descriptors for frames of audio files using librosa, a popular library for audio and music analysis written in Python.[6] After having obtained the descriptors, we cluster them using mini-batch k-means (Sculley, 2010) and quantize the descriptors into a "bag of audio words" (BoAW) (Foote, 1997) representation by comparing the MFCC descriptors to the cluster centroids. This gives us BoAW representations for each of the audio files. Auditory representations are obtained by taking the mean of the BoAW representations of the relevant audio files, and finally weighting them using positive point-wise mutual information (PPMI), a standard weighting scheme for improving vector representation quality (Bullinaria and Levy, 2007). We set $k = 300$, which equals the number of dimensions for the linguistic representations.

4.3 Multi-modal Fusion Strategies

Since multi-modal semantics relies on two or more modalities, there are several ways of combining or *fusing* linguistic and perceptual cues (Bruni et al., 2014). When computing similarity scores, for instance, we can either jointly learn the representations; learn them independently, combine (e.g. concatenate) them and compute similarity scores; or learn them independently, compute similarity scores independently and combine the scores. We call these possibilities *early*, *middle* and *late* fusion, respectively, and evaluate multi-modal models in each category.

[6] http://bmcfee.github.io/librosa.

4.3.1 Early Fusion

A good example of early fusion is the recently introduced multi-modal skip-gram model (Lazaridou et al., 2015). This model behaves like a normal skip-gram, but instead of only having a training objective for the linguistic representation, it includes an additional training objective for the visual context, which consists of an aggregated representation of images associated with the given target word. The skip-gram training objective for a sequence of training words $w_1, w_2, ..., w_T$ and a context size c is:

$$\frac{1}{T}\sum_{t=1}^{T} J_\theta(w_t)$$

where J_θ is the log-likelihood $\sum_{-c\leq j\leq c} \log p(w_{t+j}|w_t)$ and $p(w_{t+j}|w_t)$ is obtained via the softmax:

$$p(w_{t+j}|w_t) = \frac{\exp^{u_{w_{t+j}}^\top v_{w_t}}}{\sum_{w'=1}^{W} \exp^{u_{w'}^\top v_{w_t}}}$$

where u_w and v_w are the context and target vector representations for the word w respectively, and W is the vocabulary size. The objective for the multi-modal skip-gram has an additional visual objective J_{vis} (in this case a margin criterion):

$$\frac{1}{T}\sum_{t=1}^{T} J_\theta(w_t) + J_{vis}(w_t)$$

Here, we take a similar but more straightforward approach by making the auditory context a part of the initial training objective, which is possible because linguistic and auditory representations have the same dimensionality. That is, we modified word2vec to predict additional auditory contexts that have been set to the corresponding BoAW representation. We jointly learn linguistic and audio representations by taking the aggregated mean μ_w^a of the auditory vectors for a given word w, and adding this mean vector to the context:

$$\frac{1}{T}\sum_{t=1}^{T} J_\theta(w_t) + \log p(\mu_{w_t}^a|w_t)$$

The intuition is that the induced vector for the target word now has to predict an auditory vector as part of its context, as well as the linguistic ones. As an alternative, we also investigate re-

placing the mean $\mu^a_{w_t}$ with an auditory vector obtained by uniformly sampling from the set of auditory representations for the target word. We refer to these two alternatives as MMSG-MEAN and MMSG-SAMPLED, respectively. For this model, auditory BoAW representations are built for the ten thousand most frequent words in our corpus, based on 10 audio files retrieved from FreeSound for each word (or fewer when 10 are not available).

4.3.2 Middle and Late Fusion

Whereas early fusion requires a joint training objective that takes into account both modalities, middle fusion allows for individual training objectives and independent training data. Similarity between two multi-modal representations is calculated as follows:

$$sim(u, v) = g(f(u^l, u^a), f(v^l, v^a))$$

where g is some similarity function, u^l and v^l are linguistic representations, and u^a and v^a are the auditory representations. A typical formulation in multi-modal semantics for $f(x, y)$ is $\alpha\, x \| (1-\alpha) y$, where $\|$ is concatenation (see e.g. Bruni et al. (2014) and Kiela and Bottou (2014)).

Late fusion can be seen as the converse of middle fusion, in that the similarity function is computed first before the similarity scores are combined:

$$sim(u, v) = h(g(u^l, v^l), g(u^a, v^a))$$

where g is some similarity function and h is a way of combining similarities, in our case a weighted average: $h(x, y) = \frac{1}{2}(\alpha\, x + (1-\alpha) y)$; and we use $g = \frac{x \cdot y}{|x||y|}$ (cosine similarity). Since cosine similarity is the normalized dot-product, and the uni-modal representations are themselves normalized, middle and late fusion are equivalent if $\alpha = 0.5$, which we call MM. However, when $\alpha \neq 0.5$, we distinguish between the two models, calling them MM-MIDDLE and MM-LATE respectively.

5 Results

5.1 Conceptual Similarity and Relatedness

We evaluate performance by calculating the Spearman ρ_s correlation between the ranking of the concept pairs produced by the automatic similarity metric (cosine between the derived vectors) and that produced by the gold-standard similarity scores. To ensure a fair comparison, we evaluate on the common subsets for which there are representations in both modalities (see Table 2).

Modality	MEN	AMEN	SLex	ASLex
Linguistic	0.687	0.603	0.320	0.314
Auditory	0.325	0.510	0.161	0.201
MMSG-MEAN	0.612	0.537	0.274	0.266
MMSG-SAMPLED	**0.690**	0.602	**0.321**	0.314
MM	0.680	**0.662**	0.314	**0.345**

Table 3: Spearman ρ_s correlation comparison of uni-modal and multi-modal representations. The MMSG models perform early fusion, MM represents middle and late fusion with $\alpha = 0.5$ (see Section 4.3.2).

The results are reported in Table 3. We find that, while performance decreases for linguistic representations on the auditory-relevant subsets of the two datasets, performance increases for the uni-modal auditory representations on those subsets. This indicates that our auditory representations are better at judging auditory-relevant comparisons than they are at non-auditory ones, as we might expect.

For all datasets, the accuracy scores for multi-modal models are at least as high as those for the purely linguistic representations. In the case of the full datasets this difference is only marginal, which is to be expected given how few of the words in the datasets are auditory-relevant. However, the results indicate that adding auditory input even for words that are not directly auditory-relevant is not detrimental to overall performance.

In the case of the auditory-relevant subsets, we see a large increase in performance when using multi-modal representations. It is also interesting that this performance increase is found in the simple MM model, compared to the more complicated MMSG models, which seems to indicate that the latter models are still too reliant on linguistic information, which harms their performance when performing auditory-specific comparisons. The model which performs consistently well across the four datasets is MM, the middle-late fusion model with $\alpha = 0.5$.

5.2 Cross-modal Zero-shot Learning

We learn a cross-modal mapping between the linguistic and auditory spaces using partial least squares regression, taking out each concept, training on the others, and then projecting from one

Mapping	P@1	P@5	P@20	P@50
Chance	0.00	0.93	4.01	8.49
Auditory ⇒ Ling.	0.77	6.48	17.54	31.33
Ling. ⇒ Auditory	0.73	6.71	22.16	37.32

Table 4: Cross-modal zero-shot learning accuracy.

space into the other. Zero-shot performance is evaluated using the average percentage correct at N (P@N), which measures how many of the test instances were ranked within the top N highest ranked nearest neighbors. Results are shown in Table 4, with the chance baseline obtained by randomly ranking a concept's nearest neighbors. Insofar as it is possible to make a direct comparison with linguistic-visual zero-shot learning (which uses entirely different data), it appears that the current task may be more difficult: Lazaridou et al. (2014) report a P@1 of 2.4 and P@20 of 33.0 for their linguistic-visual model.

5.3 Qualitative Analysis

We also performed a small qualitative analysis of the BoAW representations for the words in MEN and SLex. As Table 5 shows, the nearest neighbors are remarkably semantically coherent. For example, the model groups together sounds produced by machines, or by water. It even finds that dinner, meal, lunch and breakfast are closely related. In contrast, nearest neighbors for the linguistic model tend to be of a more abstract nature: where we find *mouth* and *throat* as auditory neighbors for *language*, the linguistic model gives us concepts like *word* and *dictionary*; while auditory *gossip* sounds like *maids* and is something you might do in the *corridor*, it is linguistically associated with more abstract concepts like *news* and *newspaper*.

6 Parameter Tuning

There are many parameters that were left fixed in the main results that could have been adjusted to improve performance, particularly in the middle and late fusion models. It is useful to investigate some of the parameters that are likely to have an impact on performance: what the effect of the α mixing parameter is, whether a different k would have yielded better auditory representations, and whether the number and duration of the audio files from FreeSound has any effect.

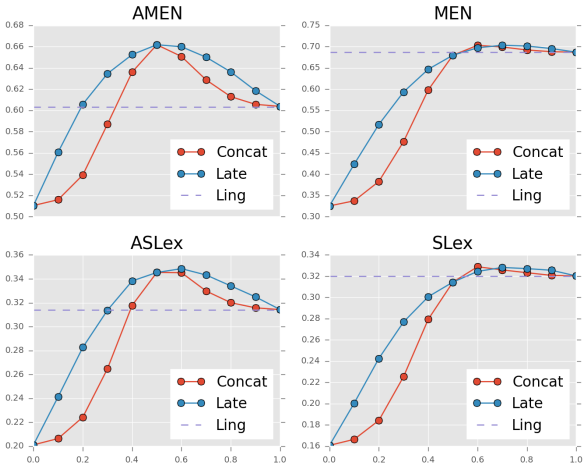

Figure 1: Performance of middle and late multimodal fusion models compared to linguistic representations on the four datasets when varying the α mixing parameter on the x-axis.

6.1 Mixing with α

The mixing parameter α plays an important role in the middle and late fusion models. We kept it fixed at 0.5 for the MM model above, but here we experiment with varying the parameter, yielding results for two different models, MM-MIDDLE and MM-LATE. The results are shown in Figure 1, where moving to the right on the x-axis uses more linguistic input and moving to the left uses more auditory input. The late model consistently outperforms the middle fusion model, which is probably because it is less susceptible to any noise in the auditory representation. Optimal performance seems to be around $\alpha = 0.6$ for both fusion strategies on all four datasets, indicating that it is better to include a little more linguistic than auditory input. It appears that any $0.5 \leq \alpha < 1$ (i.e., where we have more linguistic input but still some auditory signal), outperforms the purely linguistic representation, substantially in the case of the auditory-relevant subsets.

6.2 Number of Auditory Dimensions

We experimented with different values for the number of audio words k (i.e. the number of clusters in the k-means clustering that determines the number of "audio words"). As Figure 2 shows, the quality of the uni-modal auditory representations is highly robust to the number of dimensions. In fact, any choice of k in the range shown provides similar results across the datasets.

Auditory				Linguistic			
navy	language	gossip	dinner	navy	language	gossip	dinner
army	mouth	maid	meal	army	word	news	lunch
aviation	man	guest	lunch	military	words	newspaper	wedding
plane	father	elevator	writer	vessel	literature	cute	meal
jet	adult	danger	breakfast	sunk	dictionary	sexy	breakfast
cannon	throat	corridor	couch	ship	tongue	mirror	cocktail
monster	motor	water	dawn	monster	motor	water	dawn
orchestra	engine	stream	summer	zombie	vehicle	droplets	dusk
demon	rain	bath	child	demon	automobile	salt	sunrise
guitar	beach	river	victor	dragon	car	cold	moon
beast	boat	bathroom	morning	beast	motorcycle	sunlight	night
pilot	car	rain	garden	creatures	truck	milk	misty

Table 5: Example nearest neighbors for auditory (BoAW) representations and linguistic representations.

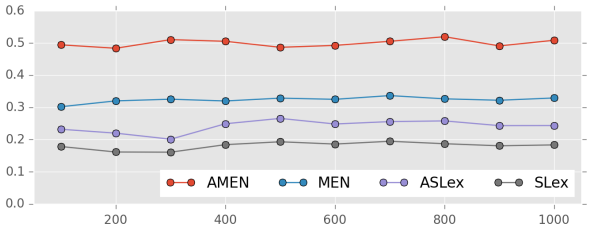

Figure 2: Performance of uni-modal auditory representations on the four datasets when varying the k parameter.

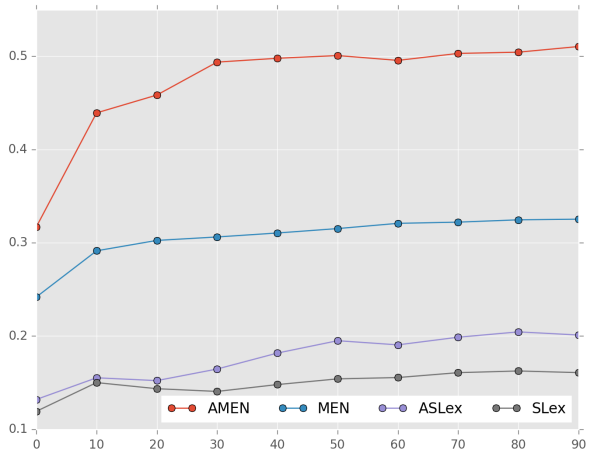

Figure 3: Performance of uni-modal auditory representations on the four datasets when varying the number of audio files per target word.

6.3 Number and Duration of Audio Files

We experimented with the number of audio files by querying FreeSound for up to 100 audio files per search word, while keeping $k = 300$. The results are shown in Figure 3. It appears that "the more the better", although performance does not increase significantly after around 40 audio files.

In order to examine the effect of audio file duration, we experimented with specifying the duration of audio files when querying the database, either taking very short (up to 5 seconds), medium length (up to 1 minute) or files of any duration. The results can be found in Figure 4, showing that performance generally increases as the files get longer (except on AMEN where a duration of 1 minute provides optimal performance).

7 Case Study: Musical Instruments

To strengthen the finding that multi-modal representations perform well on the auditory-relevant subsets of the datasets, we evaluate on an altogether different task, namely that of musical instrument classification. We used Wikipedia to collect a total of 52 instruments and divided them into 5 classes: brass, percussion, piano-based, string and woodwind instruments. For each of the instruments, we collected as many audio files from FreeSound as possible, and used the MM-MIDDLE model with parameter settings that yielded good results in the previous experiments ($k = 300$ and $\alpha = 0.6$). We then performed k-means clustering with five cluster centroids and compared results between auditory, linguistic and multi-modal, evaluating the clustering quality using the standard V-measure clustering evaluation metric (Rosenberg and Hirschberg, 2007).

This is an interesting problem because instrument classes are determined somewhat by conven-

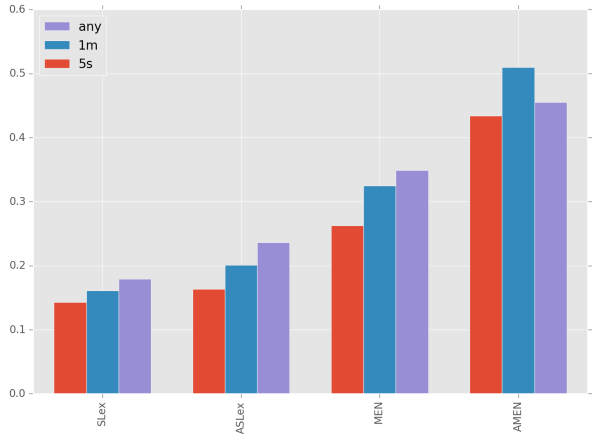

Figure 4: Performance of uni-modal auditory representations on the four datasets when varying the maximum duration.

Model	Auditory	Linguistic	MM-MIDDLE
V-measure	0.39	0.47	0.54

	Linguistic
1	baritone
2	lute, zither, xylophone, lyre, cymbals
3	piano, trombone, clarinet, cello, violin
4	castanets, tambourine, claves, maracas
5	trumpet, horn, bugle, cowbell, carillon

	Multi-modal
1	drum, claves, bongo, bass, conga
2	xylophone, glockenspiel, tambourine, cymbals
3	cello, piano, clarinet, trombone, violin,
4	chimes, bell
5	mandolin, banjo, harmonica, guitar, sitar

Table 6: V-measure performance for clustering musical instruments, together with instruments closest to cluster centroid for linguistic and multi-modal.

tion (is a saxophone a brass or a woodwind instrument?). What is more, how instruments actually sound is rarely described in detail in text, so corpus-based linguistic representations cannot take this information into account. The results are in Table 6, clearly showing that the multi-modal representation which utilizes both linguistic information and auditory input performs much better on this task than the uni-modal representations. It is interesting to observe that the linguistic representations perform better than the auditory ones: a possible explanation for this is that audio files in FreeSound are rarely samples of a single individual instrument, so if a bass is often accompanied by a drum this will affect the overall representation. The table also shows, for the 5 clusters under both models, the nearest instruments to the cluster centroids, qualitatively demonstrating the greater cluster coherence for the multi-modal model.

8 Conclusions

We have studied grounding semantic representations in raw auditory perceptual information, using a bag of audio words model to obtain auditory representations, and combining them into multi-modal representations using a variety of fusion strategies. Following previous work in multi-modal semantics, we evaluated on conceptual similarity and relatedness datasets, and on the cross-modal task of zero-shot learning. We presented a short case study showing that multi-modal representations perform much better than auditory or linguistic representations on a musical instrument clustering task. It may well be the case that the auditory modality is better suited for other evaluations, but we have chosen to follow standard evaluations in multi-modal semantics to allow for a direct comparison.

In future work, it would be interesting to investigate different sampling strategies for the early fusion joint-learning approach and to investigate more sophisticated mixing strategies for the middle and late fusion models, e.g. using the "audio dispersion" of a word to determine how much auditory input should be included in the multi-modal representation (Kiela et al., 2014). Another interesting possibility is to improve auditory representations by training a neural network classifier on the audio files and subsequently transferring the hidden representations to tasks in semantics. Lastly, now that the perceptual modalities of vision, audio and even olfaction (Kiela et al., 2015) have been investigated in the context of distributional semantics, the logical next step for future work is to explore different fusion strategies for multi-modal models that combine various sources of perceptual input into a single grounded model.

Acknowledgments

DK is supported by EPSRC grant EP/I037512/1. SC is supported by ERC Starting Grant DisCoTex (306920) and EPSRC grant EP/I037512/1. We are grateful to Xavier Serra, Frederic Font Corbera, Alessandro Lopopolo and Emiel van Miltenburg

for useful suggestions and thank the anonymous reviewers for their helpful comments.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, pages 238–247.

Shane Bergsma and Randy Goebel. 2011. Using visual information to predict lexical preference. In *Proceedings of RANLP*, pages 399–405.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artifical Intelligence Research*, 49:1–47.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting Semantic Representations from Word Co-occurrence Statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Stephen Clark. 2015. Vector Space Models of Lexical Meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics*, chapter 16. Wiley-Blackwell, Oxford.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

A. Eronen. 2003. Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. In *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications*, volume 2, pages 133–136.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Proceedings of NAACL*, pages 91–99.

Robert Fergus, Fei-Fei Li, Pietro Perona, and Andrew Zisserman. 2005. Learning object categories from Google's image search. In *Proceedings of ICCV*, pages 1816–1823.

Jonathan T Foote. 1997. Content-based retrieval of music and audio. In *Voice, Video, and Data Communications*, pages 138–147.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Proceedings of NIPS*, pages 2121–2129.

Michael S. Gazzaniga, editor. 1995. *The Cognitive Neurosciences*. MIT Press, Cambridge, MA.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42:335–346.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of EMNLP*, pages 255–265.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, pages 36–45.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*, pages 835–841.

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding semantics in olfactory perception. In *Proceedings of ACL*, pages 231–236, Beijing, China, July.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *Proceedings of ACL*, pages 1403–1414.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skipgram model. In *Proceedings of NAACL*.

Chee Wee Leong and Rada Mihalcea. 2011. Going beyond text: A hybrid image-text approach for measuring word relatedness. In *Proceedings of IJCNLP*, pages 1403–1407.

A. Lopopolo and E. van Miltenburg. 2015. Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*.

Max M. Louwerse. 2008. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 59(1):617–645.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, Arizona, USA.

D. O'Shaughnessy. 1987. *Speech communication: human and machine*. Addison-Wesley series in electrical engineering: digital signal processing. Universities Press (India) Pvt. Limited.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of EMNLP*, pages 1146–1157.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

David Sculley. 2010. Web-scale k-means clustering. In *Proceedings of WWW*, pages 1177–1178. ACM.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of EMNLP*, pages 1423–1433.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*, pages 721–732.

Josef Sivic and Andrew Zisserman. 2003. Video google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, pages 1470–1477.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of ACL*, 2:207–218.

Stanley Smith Stevens, John Volkmann, and Edwin B. Newman. 1937. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185—190.

Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. 2014. Learning words from images and speech. In *NIPS Workshop on Learning Semantics*, Montreal, Canada.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: vector space models of semantics. *Journal of Artifical Intelligence Research*, 37(1):141–188, January.

Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 319–326. ACM.