Sound-based distributional models

Alessandro Lopopolo
VU University Amsterdam
a.lopopolo@vu.nl

Emiel van Miltenburg
VU University Amsterdam
emiel.van.miltenburg@vu.nl

Abstract

Following earlier work in multimodal distributional semantics, we present the first results of our efforts to build a perceptually grounded semantic model. Rather than using images, our models are built on sound data collected from freesound.org. We compare three models: one bag-of-words model based on user-provided tags, a model based on audio features, using a 'bag-of-audio-words' approach and a model that combines the two. Our results show that the models are able to capture semantic relatedness, with the tag-based model scoring higher than the sound-based model and the combined model. However, capturing semantic relatedness is biased towards language-based models. Future work will focus on improving the sound-based model, finding ways to combine linguistic and acoustic information, and creating more reliable evaluation data.

1 Introduction

This paper explores the possibility of creating distributional semantic models (Turney et al. 2010) from a large dataset of tagged sounds.[1] Traditionally, distributional models have solely relied on word co-occurrences in large corpora. Recently, Bruni et al. (2012a) have started to combine visual features with textual ones, resulting in a performance increase for tasks focusing on the use of color terms. Sound has received relatively little attention in the distributional semantics literature, but we believe that it may be a useful source of information, especially for the representation of actions or events like *talking, shattering*, or *barking*. We propose three models: a tag-based model, a model based on bags-of-auditory words, and a model combining both kinds of information. We evaluate these models against human similarity and relatedness judgments, and show that while the tag-based model performs better than the others, the sound-based model and the combined model still correlate well with the judgment data. Based on our results, we suggest a number of future improvements. All of our code is available on GitHub.[2]

2 Three distributional models

In the following sections, we present three distributional models: a tag-based model (SoundFX-tags), a bag-of-auditory-words model (SoundFX-BoAW), and a model based on both kinds of information (SoundFX-combined). We base our models on a subset of 4,744 sounds from the Freesound database (Font et al. 2013, total 225,247 sounds) that were manually categorized as SoundFX by Font et al. (2014). We chose to focus on this subset because these sounds represent real-life events (e.g. opening and closing doors, barking dogs, and leaking faucets), rather than e.g. music. All sounds are tagged and provided with descriptions by the uploaders. General statistics of this data set are given in Table 1. Some examples of tags associated with particular sounds are {*lawn-mower, machine, motor, mower*}, and {*laundromat, laundry, machine, vibration*}. We downloaded the sounds and the metadata through the Freesound API.[3]

[1]This work was carried out in the *Understanding Language by Machines* project, supported by the Spinoza prize from the Netherlands Organisation for Scientific Research (NWO). We wish to thank three anonymous reviewers for their feedback, which has noticeably improved this paper. All remaining errors are, of course, our own.

[2]https://github.com/evanmiltenburg/soundmodels-iwcs

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

Proceedings of the 11th International Conference on Computational Semantics, pages 70–75,
London, UK, April 15-17 2015.

Total number of tags	43277	Average number of tags per sound	9
Number of different tags	4068	Average number of sounds per tag	11
Total number of sounds	4744	Average duration (seconds)	29

Table 1: General statistics about the SoundFX dataset.

2.1 A tag-based distributional model

We used Latent Semantic Analysis (Landauer & Dumais 1997) to create a tag-based distributional model (SoundFX-tags) . In our implementation, we used Scikit-learn (Pedregosa et al. 2011) to transform the tag×tag co-occurrence matrix (using TF-IDF), reduce the dimensionality of the matrix (using SVD, singular value decomposition), and to get a distance matrix using the cosine similarity measure. In our setup, we only use tags occurring more than 5 times that do not only contain digits (which are typically used to categorize files (by year, track number, etc.) rather than to describe their contents).

2.2 A sound-based distributional model

The tag-based model above provides an example of the *bag-of-words* approach, where the meaning of a word is inferred from co-occurrence data with other words. This idea has recently been extended to the domain of images; in the *bag-of-visual-words* approach, the meaning of a word is computed on the basis of visual features extracted from the images it occurs with (Bruni et al. 2012a,b, 2014). Inspired by this image-based work, researchers in the field of sound event detection have started to implement models using a *bag-of-audio-words* (BoAW) approach (Liu et al. 2010, Pancoast & Akbacak 2012). Following these authors, we set up a pipeline to create a BoAW-model. Our pipeline consists of the following stages:

1. **Preprocessing**: we convert all the sounds to WAV-format, resample them to 44100 Hz, and normalize the sounds using RMS (root mean square).
2. **Populating the feature space**: we populate a feature space by extracting acoustic descriptors from a training database of sounds. We partition each sound in partially overlapping windows of a fixed size, and compute descriptors on the basis of each window.[4] As the descriptor, we use mel-frequency cepstral coefficients (Fang et al. 2001) concatenated with log spectral energy.
3. **Building the audio vocabulary**: we cluster the descriptors using k-means. The centroids of the clusters are chosen as the audio words of our vocabulary, and thus we end up with k audio words.
4. **Creating a [sound × audio-word] matrix**: we again partition each sound in windows, and for each window we compute an acoustic descriptor. We then create a count matrix, based on which audio word is closest (using Euclidean distance) to each descriptor.
5. **Creating a [tag × audio-word] matrix**: we take the previous matrix to compute the tag×audio-word matrix. Every tag is represented by a vector that is the grand average of all the sound-vectors associated with that tag.

For our model (SoundFX-BoAW), we randomly selected 400 sounds from the set of SoundFX files. These sounds constitute the training set of the vocabulary building phase. For both the vocabulary building and the encoding phase, descriptors are extracted from each sound using windows of 250ms starting every 100ms. The BoAW model was created using 100 audio words. We transformed the values in the [tag × audio-word] matrix using Positive Local Mutual Information (plmi). The [tag × audio-word] matrix is obtained by averaging all the sound-vectors associated with each single tag (mean number of sound per tag = 39.71). Its dimensionality is reduced using SVD in order to avoid sparcity and reduce noise.

[4] The total number of windows (and thus the number of descriptors extracted) depends on the actual length of the sound.

2.3 A combined model

To combine tag information with audio word information, for this paper we propose to concatenate the TF-IDF-transformed SoundFX [tag × tag] matrix with a PLMI-transformed [tag × audio-word] matrix, and then reduce the resulting matrix using SVD. This is similar to Bruni et al.'s (2012a) approach. As an alternative to concatenation-based approaches, we can imagine creating a shared embedding model with both sound and language data (cf. Socher et al.'s (2013) work with images), but we leave the construction of such a model to future research.

3 Evaluation Procedure

We evaluated our models on the MEN Test Collection (Bruni et al. 2014) and on SimLex-999 (Hill et al. 2014). The former provides a test of semantic relatedness, i.e. how strongly two words are associated. The latter tests semantic similarity, which is a measure of how *alike* two concepts are. Hill et al. (2014) show that similarity ratings are more difficult to predict on the basis of text corpus data.

To evaluate our models, we took the tag pairs from both benchmarks, and we compared the similarity/relatedness ratings with the cosine similarity between the tags in each model. The correlation between the cosine similarity values and the actual ratings are reported in the next section. Our expectations for the performance of our models are as follows:

1. We expect the tag-based model (SoundFX-tags) to perform better on MEN than on SimLex-999, because (i) tag co-occurrence can only tell us something about which actions and entities typically occur together in a given event. But that does not tell us how alike the tags are; and (ii) language-based models typically perform better on measures of relatedness.
2. We expect the sound-based model (SoundFX-BoAW) to perform worse than SoundFX-tags on both tasks. This would parallel Bruni et al.'s results with bag-of-visual-words models.
3. We expect SoundFX-BoAW to perform better on SimLex-999 than on MEN, because the model clusters sounds (and thus tags) by their likeness.
4. We expect the combined model (SoundFX-combined) to perform better on both tasks than SoundFX-BoAW, because it is enriched with information from SoundFX-tags (which we expect to be a better approximator of human semantic judgment, see expectation 1).

In addition, we also created a tag-based model using the full Freesound database (Freesound-tags) because we were curious to see the effects of quantity versus quality (i.e. homogeneity) of tags. We were hoping that the two would balance each other out, yielding a performance on par with SoundFX-tags. In future work, we hope to be able to provide a better benchmark for sound-based distributional models. For example, MEN is based on tags from an image dataset. Clearly this is not ideal for models of sound domain, and a more balanced test collection is needed.

4 Results

Table 2 shows the results for our models. Each row shows the correlation scores on MEN and SimLex-999 for our models. We have selected the models with the optimal number of dimensions, which is why we report two models for SoundFX-tags: with 100 dimensions it produces the best score on SimLex, while with 400 dimensions it produces the best score on MEN. Our other models did not differ as much in their 'high-scores' for MEN and SimLex, which is why we do not report different versions of these. The results confirm our first two expectations, while the latter two expectations are not borne out.

SoundFX-tags scores better on MEN than on SimLex. The correlations show that our tag-based models are also more robust in capturing relatedness: optimizing on SimLex rather than on MEN is not as detrimental to the MEN-results as optimizing on MEN is to the SimLex-results. Our results also show that adding more data does not improve our model: when we use all the tags in the full Freesound database, the results are only slightly worse on the MEN task, and average on the SimLex task.

Model	Dimensions	MEN	SimLex-999
SoundFX-tags	100	0.675	**0.489**
SoundFX-tags	400	**0.689**	0.397
Freesound-tags	3000	0.643	0.426
SoundFX-BoAW	60	0.402	0.233
SoundFX-combined	1000	0.404	0.226

Table 2: Spearman correlations between our models and the MEN and Simlex-999 datasets.

SoundFX-tags scores better on both evaluation sets than SoundFX-BoAW. As noted in the previous section, this parallels Bruni et al.'s (2012a) results with bag-of-visual-words models. To further compare SoundFX-tags (optimized on MEN) with SoundFX-BoAW, we computed the pairwise distance for all combinations of tags in both models. A Spearman rank correlation between these sets of distances returns a value of 0.23. This means that both models are not strongly correlated with each other. In other words: they encode different information.

Counter to our expectations, SoundFX-BoAW does *not* perform better on SimLex than on MEN. Why is this? We believe that the reason lies in the tags provided with the sounds. For a good performance on SimLex, the tags should be close descriptions of the sounds, so that we know how similar two tags are on the basis of the similarity between the sounds they denote. We know that sounds seem to be good at capturing *events* (e.g. walking), but it is often hard or even impossible to recognize the objects present in those events (e.g. shoes vs. sandals). Ideally for our purposes, the sounds in the Freesound database would only be tagged with event-descriptions or with low-level descriptions of the sounds (*walking, sniffing, barking, scratching, beeping*). However, the tags are provided by the producers of the sounds. This has the result that the tags are typically very high-level descriptions, including all the objects involved in the production of the sound. This is detrimental to the performance on SimLex. Meanwhile, because on average there are so many tags per sound, there are many tags that get clustered together based on co-occurrence. This boosts the performance on the MEN-task.

Also counter to our expectations, Table 2 shows that the combined model is equivalent (rather than superior) to the sound-based model, i.e. combining the tag data with the audio-word data does not improve performance. Following the suggestions of two reviewers, we experimented with normalizing the matrices before concatenation. We tried two methods: (i) MAXDIVIDE: Divide all values in each matrix by their respective maximum value in that matrix; (ii) LOGMAXDIVIDE: to prevent extreme values in the BoAW-matrix from marginalizing the other values, we first took the log of all values in the matrix before dividing all values in both matrices by their respective maximum values. Table 3 provides a comparison of all concatenation methods.

Model	Dimensions	MEN	SimLex-999
SoundFX-combined	1000	0.404	0.226
SoundFX-MaxDivide	150	**0.688**	0.450
SoundFX-MaxDivide	100	0.678	**0.493**
SoundFX-LogMaxDivide	150	0.442	0.232
SoundFX-LogMaxDivide	100	0.435	0.239

Table 3: Spearman correlations between different concatenation models and the MEN and Simlex-999 datasets.

We conclude that the low performance of the old concatenation model is due to the fact that the relatively high values in the BoAW matrix (between 0 and 106312) had a marginalizing effect on the low values in the tag matrix (between 0 and 1), dominating the SVD. The LOGMAXDIVIDE method delivers models that are slightly better than the BoAW model and the old concatenation model, but worse than the others. Normalization using MAXDIVIDE delivers a competitive model on the MEN dataset, and the best model overall on SimLex, beating SoundFX-tags-400 by a margin of 0.004. With such a small

difference, we should be wary of concluding that audio information (rather than chance) helped improve SimLex performance (as we predicted). More research on the possible contributions of audio data to semantic models is required.

5 Discussion and conclusion

We presented the first (to our knowledge) attempt to create a sound-based semantic space based on user-generated and -tagged sounds. The database used for our models is less controlled and much noisier as compared to other audio datasets used to test acoustic event recognition pipelines [5] or for studying auditory perception (Cotton & Ellis 2011, Capilla et al. 2013). The Freesound database (Font et al. 2013) provides a huge collection of sounds recorded by different human users, in different conditions and using different equipments. Moreover, all the linguistic metadata (tags, titles, and descriptions) are provided by the users who recorded and uploaded the sounds without any overt supervision or strict guidelines.[6] Moreover, the number of sounds per tag (i.e. the number of sounds used to encode each tag in the SoundFX-BoAW model) varies between 6 and 2050, with an average number of sound per tag equal to 39.71. What might seem like limitations at first sight, might turn out to be strong points in favor of our analyses. The strong correlations of both tag-based and sound-based families of models with the MEN and SimLex benchmarks proves that, even with uncontrolled data, the models are still able to capture semantic regularities in the auditory domain.

The sound-based model implemented the BoAW-paradigm using mel frequency cepstral coefficients sampled in partially overlapping windows over the sound signals. This captures short-term distribution of energy over the sound spectrum. BoAW encoding thus capture distributions of features across sounds, but completely ignores the development of the sound over a span of time larger than the sampling time windows. To combat this issue, we plan to assess other features or combination of features in the future (see Breebaart & McKinney 2004 for a list). Other possible improvements are using a different clustering algorithm for vocabulary creation, and using different sound encoding techniques. We plan to look at Gaussian Mixture Models and Fisher encoding (also used successfully by Bruni et al. (2012b, 2014) for their image-based models). Nonetheless, we are planning to overtake the feature selection step by exploring the possibility of using unsupervised methods for learning feature representations provided by deep learning algorithms (Bengio 2009, Mohamed et al. 2009, Hinton et al. 2012).

The evidently better performance of SoundFX-tags as compared to SoundFX-BoAW in the quantitative evaluation may simply be due to the nature of the benchmarks that we used. MEN and SimLex are both linguistic tasks, and so they may very well be biased towards language-based models. Furthermore, as mentioned in Section 3, MEN was based on a collection of image tags. Thus we may question the reliability of MEN as a test for sound-based models, which may very well stand out once sound-related tags are considered as well. In the short term, we plan to collect a controlled set of such tags, along with human similarity ratings for both the tags as well as the sounds labeled with those tags. A longer-term goal is to create other kinds of benchmarks that are better equipped to test the quality of perceptually grounded models.

Moreover, we are aware of the limitations of the simple concatenation technique employed to obtain SoundFX-combined. In order to try to overcome this, we are planning to explore deep-learning techniques to learn shared text-sound embeddings similar to what has been already proposed by Socher et al. (2013) and Ngiam et al. (2011). Our results are an encouraging first step to incorporate sound information in distributional semantic models. We hope that this development will parallel the success of image-based models, and that all kinds of perceptually grounded models may ultimately be unified in a single model that captures the semantics of human experience.

[5] `http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/description.html`

[6] Note that descriptions are moderated, and users may receive a warning that their sounds have a "bad description". There are some description guidelines, which can be found at: `http://www.freesound.org/help/faq/#sounds-2`. However, this does not mean that all sounds have a clear and uniform description.

References

Bengio, Yoshua. 2009. Learning deep architectures for ai. *Found. Trends Mach. Learn.* 2(1). 1–127. doi:10.1561/2200000006. http://dx.doi.org/10.1561/2200000006.

Breebaart, J. & M. McKinney. 2004. Features for audio classification .

Bruni, Elia, Gemma Boleda, Marco Baroni & Nam-Khanh Tran. 2012a. Distributional semantics in technicolor. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1*, 136–145. Association for Computational Linguistics.

Bruni, Elia, Nam Khanh Tran & Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 1. 1–47.

Bruni, Elia, Jasper Uijlings, Marco Baroni & Nicu Sebe. 2012b. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of acm multimedia*, 1219–1228. Nara, Japan.

Capilla, Almudena, Pascal Belin & Joachim Gross. 2013. The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cerebral cortex (New York, N.Y. : 1991)* 23(6). 1388–95. doi:10.1093/cercor/bhs119.

Cotton, Courtenay V. & Daniel P. W. Ellis. 2011. Spectral vs. spectro-temporal features for acoustic event detection. In *Ieee workshop on applications of signal processing to audio and acoustics, waspaa 2011, new paltz, ny, usa, october 16-19, 2011*, 69–72.

Fang, Zheng, Zhang Guoliang & Song Zhanjiang. 2001. Comparison of different implementations of mfcc. *J. Comput. Sci. Technol.* 16(6). 582–589. doi:10.1007/BF02943243. http://dx.doi.org/10.1007/BF02943243.

Font, Frederic, Gerard Roma & Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st acm international conference on multimedia*, 411–412. ACM.

Font, Frederic, Joan Serrà & Xavier Serra. 2014. Audio clip classification using social tags and the effect of tag expansion. In *Audio engineering society conference: 53rd international conference: Semantic audio*, Audio Engineering Society.

Hill, Felix, Roi Reichart & Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456* .

Hinton, Geoffrey E., Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath & Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* 29(6). 82–97.

Landauer, Thomas K & Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104(2). 211.

Liu, Yang, Wanlei Zhao, Chong-Wah Ngo, Changsheng Xu & Hanqing Lu. 2010. Coherent bag-of audio words model for efficient large-scale video copy detection. In Shipeng Li, Xinbo Gao & Nicu Sebe (eds.), *Civr*, 89–96. ACM.

Mohamed, Abdel-rahman, George E. Dahl & Geoffrey E. Hinton. 2009. Deep belief networks for phone recognition. In *Nips workshop on deep learning for speech recognition and related applications*, .

Ngiam, Jiquan, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee & Andrew Y Ng. 2011. Multimodal Deep Learning. *Proceedings of The 28th International Conference on Machine Learning (ICML)* 689–696.

Pancoast, Stephanie & Murat Akbacak. 2012. Bag-of-audio-words approach for multimedia event classification. In *Interspeech*, ISCA.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Socher, Richard, Milind Ganjoo, Christopher D Manning & Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, 935–943.

Turney, Peter D, Patrick Pantel et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37(1). 141–188.