Audio-based Distributional Semantic Models for Music Auto-tagging and Similarity Measurement

Giannis Karamanolakis*, Elias Iosif*, Athanasia Zlatintsi*, Aggelos Pikrakis† and Alexandros Potamianos*

*School of Electrical and Computer Engineering, National Technical University of Athens, Greece
†Department of Informatics, University of Piraeus, Greece
giannis.karamanolakis@gmail.com, iosife@central.ntua.gr, nzlat@cs.ntua.gr, pikrakis@unipi.gr, potam@central.ntua.gr

Abstract—The recent development of Audio-based Distributional Semantic Models (ADSMs) enables the computation of audio and lexical vector representations in a joint acoustic-semantic space. In this work, these joint representations are applied to the problem of automatic tag generation. The predicted tags together with their corresponding acoustic representation are exploited for the construction of acoustic-semantic clip embeddings. The proposed algorithms are evaluated on the task of similarity measurement between music clips. Acoustic-semantic models are shown to outperform the state-of-the-art for this task and produce high quality tags for audio/music clips.

I. INTRODUCTION

Semantic information in the form of metadata, e.g., tags, has been valuable in enhancing the performance for many music processing tasks [1], [2], [3]. Metadata typically comes in two forms: free text associated with a music or audio clip or tags (list of words or phrases) that describe the clip. Tags are often preferred over (web-mined) text, because they give a direct description of the song, e.g., genre or instruments, while the latter is inherently noisy; only a part of the text is musically relevant. The automatic annotation of clips (auto-tagging) is becoming vital and finds numerous applications including efficient music indexing and retrieval.

A variety of auto-tagging methods have been proposed in the literature. In [4], the similarity between artists is exploited for the prediction of the most descriptive tags for clips. In [5], language models were computed using Restricted Boltzmann Machines, while in [6], the combination of audio features is proposed within a block-level framework. The use of semantic tags for music similarity measurement is proposed in [7], where each song is represented by a Semantic Multinomial Distribution over a vocabulary of tags.

Music similarity is at the core of query-by-example, where the user gives a musical piece as a query and the system returns a ranked list of recommendations. Content-based similarity can be exploited by collaborative filtering algorithms [8], [9] especially where there is lack of collaborative filtering data, a.k.a. "cold start" problem. This improves the efficiency of music recommendation and playlist generation, two important tasks in Music Information Retrieval (MIR). Music similarity estimation can be formulated as the problem of finding an appropriate embedding of a music clip with respect to a distance metric. Many approaches use machine learning techniques in order to learn the distance metric that best approximates absolute user ratings [10], [11], [12] or relative user ratings

from different datasets [13]. Furthermore, research efforts include the collection of similarity scores that can be used as groundtruth information [14], [15] and the investigation of different evaluation techniques [16] for music similarity.

Distributional Semantic Models (DSMs) [17] is a popular method for automatically constructing semantic representations from text. Despite their success in various semantic tasks (e.g., semantic classification and computation of semantic similarity), the DSMs have been criticized as "disembodied", since they rely solely on linguistic information without being grounded in perception and action. The disconnection of natural language from the physical world, also referred as the symbol grounding problem [18], is alleviated via the integration of multiple modalities [19], [20]. The development of audio-based DSMs (ADSMs) was proposed in [21] for the representation of words based on their acoustic properties, while in [22] an extension was presented using combinations of auditory and linguistic features. A recently proposed approach dealt with the fusion of (different) acoustic features according to the nature of sounds (music, speech, other) [23].

In this work, the ADSM described in [23] is adopted for the computation of audio and lexical vector representations in a joint acoustic-semantic space. These 'bag-of-audio-words' representations are used for the automatic annotation of music clips. Then, the predicted tags and the acoustic features are exploited for the construction of acoustic-semantic clip embeddings. The proposed algorithms are evaluated on the task of music similarity measurement between clips taken from the MagnaTagATune dataset [14].

This paper is organized as follows. In Section II, various methods are described for the construction of clip representations. In Section III, the experimental dataset and procedure are described in detail, while in Section IV, the evaluation results of the proposed methods are reported and compared with the literature.

II. SYSTEM DESCRIPTION

For a dataset consisting of audio clips and the corresponding annotation (tags), the procedure for the creation of clip representations consists of a series of steps described below.

A. Creation of audio-word vocabulary

In order to train an audio-word vocabulary, the clips are partitioned into partially overlapping windows and a feature
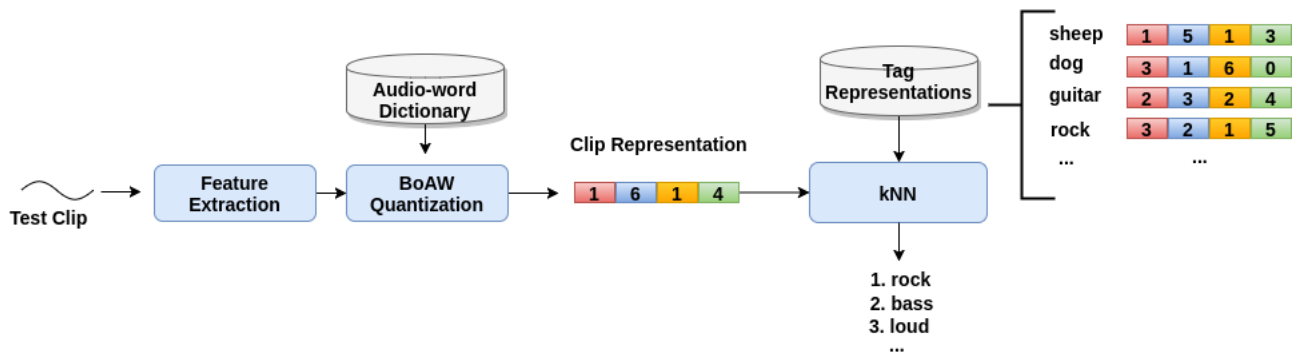
Fig. 1: Auto-tagging system overview.

Clip id	Groundtruth Tags	Predicted Tags (N=5)
3843	**indian**, **sitar**	**sitar**, **indian**, eastern, india, oriental
9531	rock, **heavy**, **heavy metal**, **loud**, fast, hard rock, **metal**	hard, **loud**, **heavy**, **heavy metal**, **metal**
13526	bass, **drums**, drum, **funky**, **reggae**	**funky**, beat, **drums**, **reggae**, funk
15380	**classical**, **solo**, **cello**, **violin**, strings	**cello**, viola, **violin**, **solo**, **classical**
19920	-	orchestra, violins, flutes, fiddle, violin
21725	choir, **choral**, **men**, man	monks, chant, chanting, **men**, **choral**
29231	**acoustic**, **guitar**	classical guitar, **guitar**, **acoustic**, lute, spanish
43390	**rock**, loud, **pop**, vocals, **male vocals**	**male vocals**, **pop**, male vocal, male singer, **rock**
48010	silence	low, soft, no singing, quiet, wind
57081	**piano**	piano solo, **piano**, classic, solo, classical

TABLE I: Examples of auto-tagging outputs for the MagnaTagATune dataset. The predicted tags are ordered according to their similarity with the respective audio clip (in descending order).

vector is extracted from every window. Therefore, every clip is represented by a set of vectors depending on its length. Next, all vectors are clustered by applying the k-means algorithm and the k centroids of the returned clusters are considered as the audio-words of the vocabulary.

B. Audio clip representations

As presented in [23], audio clips are represented as a mixture of audio-words (bag-of-audio-words approach). For each window $\vec{o_t}$ of a clip, a feature vector $\vec{x_t} \in R^d$ is computed (where d is the dimensionality of the feature space) and associated with the audio-word vocabulary. The association is performed by encoding the $\vec{x_t}$ vector as a k-dimensional vector $\vec{e_t}$:

$$\vec{e_t} = (w_1, w_2, ..., w_k), \qquad (1)$$

where $w_i \in [0, 1]$ and $\sum_{i=1}^{k} w_i = 1$. The weight w_i denotes the contribution of the i-th audio-word to the window representation and is computed according to the similarity score between the audio-word and the feature vector $\vec{x_t}$ (see [23] for details). Finally, the clip representation, $\vec{r_c}$, is obtained by averaging the vectors computed for the respective windows. Given a collection consisting of M clips, this process results in a $M \times k$ matrix. The space consisting of bag-of-audio-words clip representations, will be referred as AUDIO space.

C. Tag representations via the ADSM

As described in [23], an ADSM is constructed that provides bag-of-audio-words embeddings for tags, based on their asso-ciation with the clips. In particular, the representation $\vec{r_j}$ of a tag j is computed by averaging the representations of clips that have this tag in their descriptions. For a collection of audio clips with T (unique) tags this results in a $T \times k$ matrix. Then, the Positive Pointwise Mutual Information (PPMI) weighting is applied to the matrix [24], [25], [26].

D. Clip annotation (auto-tagging)

The bag-of-audio-words representations of tags provide a straightforward way for the automatic annotation (auto-tagging) of an audio clip c. First, the clip gets a represen-tation, $\vec{r_c}$ in the AUDIO Space. Then, the cosine similarity is computed between $\vec{r_c}$ and the representation $\vec{r_j}$ of each tag j:

$$s_{cj} = \frac{\vec{r_c} \cdot \vec{r_j}}{|\vec{r_c}| \cdot |\vec{r_j}|}. \qquad (2)$$

The N tags that best describe the clip c are those correspond-ing to the N highest similarity scores s_{cj} (k-NN classification). An example of the proposed auto-tagging algorithm is shown in Figure 1.

Table I includes some examples selected from the MagnaTa-gATune dataset (see Section III-A) for which the groundtruth labels are compared with the $N = 5$ automatically predicted tags. It is observed that many tags appear both in the predicted and the groundtruth labels, while other tags have very similar meaning (e.g., 'silence' and 'quiet'). Moreover, quite descrip-tive tags are returned for clips that have no annotations (e.g., clip 19920).
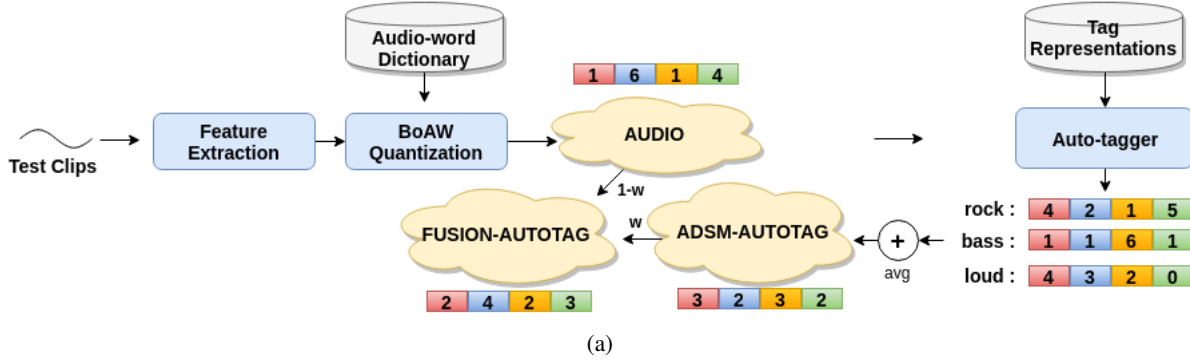
Fig. 2: Overview of music similarity estimation using AUDIO, ADSM-AUTOTAG, FUSION-AUTOTAG.

E. Semantic representations of clips

The ADSM described in Section II-C is exploited for the representation of audio clips via their association with tags. Specifically, the embedding of a clip is obtained by averaging the representations of tags that are semantically related with the clip. The determination of clip-tag associations can be performed either using already provided metadata, or via the proposed auto-tagging scheme. In the latter case, the clip gets a representation (in the ADSM-AUTOTAG space), as follows:

$$\vec{r_c'} = \frac{1}{N} \sum_{i=1}^{N} \vec{r_i}, \tag{3}$$

where $\vec{r_i}$ is the representation of the i-th tag, while N denotes the number of tags returned by the auto-tagger.

F. Fusion of audio and semantic representations

The similarity score between two clips in the AUDIO space is computed as the cosine similarity of their acoustic representations. Similarly, the similarity between clips in the ADSM space is computed as the cosine similarity of their respective semantic representations. However, the characterization of clips only via their semantic representations in some cases may lead to inaccurate estimates. For example, if two clips are annotated with the same labels, they will have exactly the same ADSM representations although they sound different. This problem can be alleviated with the fusion of acoustic and semantic information. The clip embedding $\vec{r_c}$ in the AUDIO space can be combined with the embedding from the ADSM (or ADSM-AUTOTAG) space, $\vec{r_c'}$, via a weighted average scheme:

$$\vec{r_c''} = w\vec{r_c'} + (1-w)\vec{r_c}, \tag{4}$$

or via a weighted concatenation:

$$\vec{r_c''} = w\vec{r_c'} \oplus (1-w)\vec{r_c}, \tag{5}$$

where $\oplus$ stands for the vector-concatenate operator and w is a real-valued number indicating the relative importance of the fused representations. This method results in clip embeddings in the FUSION (or FUSION-AUTOTAG) space. The various methods for the representation of clips are graphically represented along with an auto-tagging example in Fig 2(a).

III. EXPERIMENTAL DATASET AND PROCEDURE

The proposed algorithms are evaluated for the similarity computation between songs. For this purpose, the MagnaTagATune dataset is used.

A. The MagnaTagATune dataset

The MagnaTagATune dataset contains 25,863 30-second audio clips (provided by the Magnatune[1] label) and 188 tags. In addition, similarity data have been collected from the TagATune game [27]. In a bonus part of the game, the user listens to three songs and gives a vote to the song that sounds as the most dissimilar when compared with the other two songs (often called as odd one out game). The similarity data for every triplet of clips is stored in form of a triplet representing the histogram of votes i.e., the clip associated with the maximum value is the most irrelevant clip. In total, 533 triplets are derived and every triplet of song ids is saved in the form (a, b, c) where c is the outlier. This explicitly means that $d(a,b) < d(a,c)$ (constraint (a,b,c)) and $d(b,a) < d(b,c)$ (constraint (b,a,c)), where $d()$ is the perceptual distance between two clips.

Due to the objective opinion of each user, there are constraints that contradict each other. Hence, a method was proposed in [11] in order to deal with the inconsistent constraints. In [10], the 860 remaining constraints derived from [11] were split into non-overlapping training and test sets of 774 and 86 constraints, respectively enabling 10-fold cross-validation. These constraints are published serving as a common evaluation benchmark. The described benchmark is used in this work.

B. Data preprocessing and feature extraction

All audio clips are converted to WAV format and resampled at 22.05 kHz. For each clip, a feature vector is extracted from windows of 250 ms with a step of 100 ms. Here, two different feature vectors are extracted: the first (EchoNest) is obtained by the EchoNest API 1.0[2]. Specifically, 24 features per audio frame have been kept, consisting of 12 chroma features and

[1]http://magnatune.com/
[2]http://developer.echonest.com/

Literature Method	EchoNest Features
Euclidean [13]	0.598
RITML [13]	0.711
SVM [10]	**0.712**
MLR [10]	0.689

TABLE II: Accuracy of methods reported in the literature [10], [13].

Proposed Method	EchoNest		MFCCdd	
	k=300	l=10	k=300	l=10
AUDIO	0.613	0.644	0.636	0.646
ADSM	0.705	0.719	0.717	0.720
FUSION	0.720	**0.731**	0.681	0.684
ADSM-AUTOTAG	0.705	0.705	0.693	0.696
FUSION-AUTOTAG	0.705	0.709	0.662	0.672

TABLE III: Accuracy of proposed methods for EchoNest and MFC-Cdd features (for FUSION, w=0.9, while for AUTOTAG, N=20).
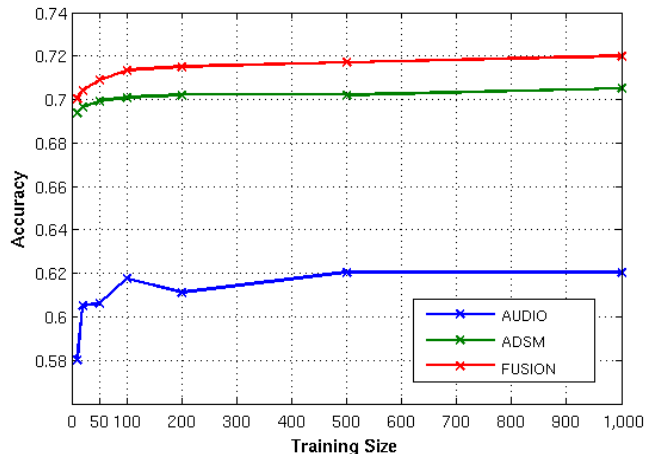

Fig. 3: Accuracy w.r.t. different training sizes, i.e., number of MagnaTagATune clips used to train the audio-word vocabulary. Results are shown for EchoNest features using $k = 300$.

12 timbre features. The chroma features describe the relative dominance of every pitch in the chromatic scale and are normalized to $[0, 1]$. The timbre features correspond to the coefficients of 12 basis functions which represent the texture of sound. The second type of feature vectors (MFCCdd) consists of the Mel Frequency Cepstral Coefficients (MFCCs) (concatenated with spectral energy), their 1st and 2nd order derivatives, resulting in a vector of 39 coefficients. In both cases, the feature vectors are normalized by their mean and standard deviation values (Z-normalization).

C. Experimental procedure

For each step of the 10-fold cross validation, the audio-word vocabulary and the ADSM are built using the training clips[3]. The representations are then computed for the test clips and the similarity scores are obtained, as described in Section II-F. The following methods are evaluated:

- **AUDIO:** only the acoustic features are used (Section II-B).
- **ADSM:** clip representations are obtained via the provided tags (Section II-E).
- **ADSM-AUTOTAG:** same with ADSM method, but all clip representations (even for clips that have labels) are derived via auto-tagging (Section II-D). Here, $N = 20$ tags[4] are predicted for every clip.
- **FUSION:** fusion of AUDIO and ADSM representations (Section II-F, for $w = 0.9$).
- **FUSION-AUTOTAG**: fusion of AUDIO and ADSM-AUTO-TAG representations, for w=0.9 and N=20.

The number of audio-words (i.e., the number of clusters) is fixed to $k = 300$. In addition, dimensionality reduction via Singular Value Decomposition (l denotes the number of dimensions) was optionally performed on the matrix where the rows correspond to different audio clips and the columns to their representations.

The adopted evaluation metric is the accuracy of each method, which is defined as the percentage of total test constraints (see Section III-A) that are satisfied. The experimental procedure is followed for each of the 10 folds and the final score is computed as the average of the accuracy scores. This procedure is repeated 10 times.

IV. EVALUATION RESULTS

Table II includes the state-of-the-art performance (see [13] for a brief overview), while the accuracy of the proposed

methods is presented in Table III. In addition to clip representations, where k=300, the performance is reported with respect to l=10 dimensions where the best accuracy is achieved for most methods. The FUSION method[5] yields the best accuracy score: 0.731, which is higher than the best score reported in the literature (0.712 achieved by SVM [10]). Moreover, the exploitation of semantic information via the ADSM method boosts the performance compared with the AUDIO method (up to 12.7% relative improvement). Regarding the ADSM-AUTOTAG, the auto-tagging algorithm is applied for all the test clips, which were not used for the training representations. However, the ADSM-AUTOTAG method achieves comparable performance with the ADSM method, where the dataset labels are used.

The quality of the auto-tagger's predictions was also confirmed after the manual inspection of the predicted tags for the examples of Table 1. Hence, the auto-tagger's predictions can be exploited for the annotation of clips without tags or for the enrichment of provided annotations. Despite the fact that the training similarity data (here, in the form of constraints) are not used, the reported results exceed the state-of-the-art. As such, the proposed unsupervised algorithm can be applied in any dataset without the need of manually created similarity data.

---

[3]For computational efficiency, 1000 clips are randomly selected from the training set for the audio-word vocabulary.

[4]N was set to 20 based on experiments conducted using held-out data.

[5]The performance of the weighted concatenation (see (5)) was found to be comparable with the performance of the weighted average (see (4)), so the respective results are not reported here.

In Fig 3, the accuracy of the AUDIO, ADSM and FUSION methods is shown as a function of the number of clips, which were used for the construction of the audio-word vocabulary. Interestingly, even a small number of clips (50-200) is sufficient for the creation of a vocabulary of audio-words, achieving 0.717 accuracy (for FUSION method and 200 training clips). In addition, it appears that the ADSM and FUSION methods are more robust with respect to data sparsity compared to the AUDIO method.

V. Conclusions

In this work, an audio-based DSM is exploited for creating lexical representations and the clip representations are derived via an auto-tagging scheme. The evaluation is performed on the task of music similarity measurement using the MagnaTagATune dataset. Although the similarity ratings provided with the dataset are not considered in the proposed algorithms, they exceed state-of-the-art methods, that use them in order to train a similarity metric. Moreover, very few data, e.g., 50 clips, are adequate to train the vocabulary, which is used for the creation of bag-of-audio-words representations. Therefore, this unsupervised algorithm can be applied in every dataset where neither clip annotations or similarity data are necessary. Regarding future work, we aim to investigate the fusion of acoustic, semantic and visual modalities and test the proposed algorithm on multimedia data.

Acknowledgment

References

[1] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 383–395, 2009.

[2] Y.-C. Lin, Y.-H. Yang, and H. H. Chen, "Exploiting online music tags for music emotion classification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 7, no. 1, p. 26, 2011.

[3] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.

[4] J. H. Kim, B. Tomasik, and D. Turnbull, "Using artist similarity to propagate semantic information." in *International Society of Music Information Retrieval*, vol. 9, 2009, pp. 375–380.

[5] M. I. Mandel, R. Pascanu, D. Eck, Y. Bengio, L. M. Aiello, R. Schifanella, and F. Menczer, "Contextual tag inference," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 7, no. 1, p. 32, 2011.

[6] K. Seyerlehner, M. Schedl, P. Knees, and R. Sonnleitner, "Draft: A refined block-level feature set for classification, similarity and tag prediction," *Dept. of Computational Perception, Johannes Kepler University, Linz, Austria*, 2011.

[7] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[8] B. McFee, L. Barrington, and G. Lanckriet, "Learning content similarity for music recommendation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2207–2218, 2012.

[9] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Advances in neural information processing systems*, 2013, pp. 2643–2651.

[10] D. Wolff, S. Stober, A. Nürnberger, and T. Weyde, "A systematic comparison of music similarity adaptation approaches," in *International Society of Music Information Retrieval*, 2012, pp. 103–108.

[11] S. Stober and A. Nürnberger, "An experimental comparison of similarity adaptation approaches," in *International Workshop on Adaptive Multimedia Retrieval*, 2011, pp. 96–113.

[12] M. Slaney, K. Weinberger, and W. White, "Learning a metric for music similarity," in *International Symposium on Music Information Retrieval (ISMIR)*, 2008.

[13] D. Wolff, A. MacFarlane, and T. Weyde, "Comparative music similarity modelling using transfer learning across user groups," *Proceedings of the 16th ISMIR Conference, 2015*.

[14] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging." in *International Society of Music Information Retrieval*, 2009, pp. 387–392.

[15] D. P. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence, "The quest for ground truth in musical artist similarity." in *International Society of Music Information Retrieval*. Paris, France, 2002.

[16] B. Logan, D. P. Ellis, and A. Berenzweig, "Toward evaluation techniques for music similarity," *The MIR/MDL Evaluation Project White Paper Collection*, vol. 3, pp. 81–85, 2003.

[17] M. Baroni and A. Lenci, "Distributional memory: A general framework for corpus-based semantics," *Computational Linguistics*, vol. 36, no. 4, pp. 673–721, 2010.

[18] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[19] E. Bruni, G. B. Tran, and M. Baroni, "Distributional semantics from text and images," in *Proceedings of the workshop on geometrical models of natural language semantics*, 2011, pp. 22–32.

[20] E. Iosif and A. Potamianos, "Crossmodal network-based distributional semantic models," *10th Language Resources and Evaluation Conference*, 2016.

[21] A. Lopopolo and E. van Miltenburg, "Sound-based distributional models," in *Proceedings of the 11th International Conference on Computational Semantics*, 2015, pp. 70–75.

[22] D. Kiela and S. Clark, "Multi- and cross-modal semantics beyond vision: Grounding in auditory perception," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2461–2470.

[23] G. Karamanolakis, E. Iosif, A. Zlatintsi, A. Pikrakis, and A. Potamianos, "Audio-based distributional representations of meaning using a fusion of feature encodings," *Interspeech 2016*, pp. 3658–3662, 2016.

[24] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study," *Behavior research methods*, vol. 39, no. 3, pp. 510–526, 2007.

[25] ——, "Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd," *Behavior research methods*, vol. 44, no. 3, pp. 890–907, 2012.

[26] D. Kiela and S. Clark, "A systematic study of semantic vector space model parameters," in *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL*, 2014, pp. 21–30.

[27] E. Law and L. Von Ahn, "Input-agreement: a new mechanism for collecting data using human computation games," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 1197–1206.