INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Audio-based Distributional Representations of Meaning Using a Fusion of Feature Encodings

Giannis Karamanolakis[1], *Elias Iosif*[1], *Athanasia Zlatintsi*[1],
Aggelos Pikrakis[2], *Alexandros Potamianos*[1]

[1]School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[2]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

Recently a "Bag-of-Audio-Words" approach was proposed [1] for the combination of lexical features with audio clips in a multimodal semantic representation, i.e., an Audio Distributional Semantic Model (ADSM). An important step towards the creation of ADSMs is the estimation of the semantic distance between clips in the acoustic space, which is especially challenging given the diversity of audio collections. In this work, we investigate the use of different feature encodings in order to address this challenge following a two-step approach. First, an audio clip is categorized with respect to three classes, namely, music, speech and other. Next, the feature encodings are fused according to the posterior probabilities estimated in the previous step. Using a collection of audio clips annotated with tags we derive a mapping between words and audio clips. Based on this mapping and the proposed audio semantic distance, we construct an ADSM model in order to compute the distance between words (lexical semantic similarity task). The proposed model is shown to significantly outperform (23.6% relative improvement in correlation coefficient) the state-of-the-art results reported in the literature.

Index Terms: Bag-of-audio-words, audio representations, feature space fusion, lexical semantic similarity.

1. Introduction

The creation of lexical descriptions for auditory perceptual experiences is a challenging task [2]. For the human perceptual system the auditory stimuli convey mainly low-level information [3] while lexico-semantic descriptions are processed and modeled via high-level cognitive processing [4]. In [5], acoustic and semantic models are defined as multidimensional spaces aimed for the representation of sounds and words, respectively. Obtaining crossmodal representations is desirable, especially for applications such as information retrieval systems [4]. We argue that a broader application of acoustic-semantic maps and representations could be the development of cognitive models motivated by evidence that multiple modalities contribute to the acquisition and representation of semantic knowledge [6], [7].

Query-by-example (QBE) constitutes one of most widely-used techniques in the framework of music information retrieval (MIR). Audio similarity is at the core of QBE based on features extracted from the audio signal, e.g., Mel-Frequency Cepstral Coefficients (MFCCs). For the case of similarity computation between music clips, numerous music-related features have been exploited such as timbre and rhythm. For example, in [8] distributions of timbre-related features were used for computing the Kullback-Leibler distance between songs. In [9], spectral features and clustering algorithms were investigated for the task of music genre classification. For most MIR applications, the computation of similarities relying solely on acoustic distances was observed to exhibit a number of undesirable properties, e.g., some clips were found to serve as hubs exhibiting high similarities with the majority of the collection clips [10]. It was suggested that this problem can be alleviated via the use of other knowledge sources such as textual content (tags) that is associated with the audio clips. Towards this direction, the joint modelling of the audio content and textual artist names was proposed in [11] with application to a variety of tasks including artist/song/tag prediction and the identification of similar songs/artists.

The development of crossmodal Distributional Semantic Models (DSMs) constitutes a recent research effort, focusing mainly on the fusion of textual and visual features [12]. The exploitation of audio-based features for building DSMs is a less-researched area, which is essential for the greater vision of truly multimodal DSMs. Regarding audio, a step towards this direction was the proposal of the "bag-of-audio-words" (BoAW) model motivated by the text-based "bag-of-words" model. One of the first applications based on the BoAW model was presented in [13] for the task of content-based video copy detection followed by [14] for multimedia event detection. The creation of audio-based DSMs was proposed in [1] and extended in [15] by combining both auditory and linguistic features.

In this work, we adopt the baseline BoAW model and we propose an extension via the fusion of feature spaces. An audio clip is categorized with respect to three classes, namely, music, speech and other. Next, the feature spaces are fused according to their posterior probabilities. Using a collection of audio clips annotated with tags we derive a mapping between words and audio clips. The proposed fusion is evaluated for the task of semantic similarity computation between words, outperforming the state-of-the-art results reported in the literature.

2. System description

In this section, we describe the main system components for constructing acoustic-semantic maps and representations. Each clip is associated with metadata consisting of textual tags aimed for the description of the audio content, e.g., audio clip "172712.wav" is described by the following set of tags: "animal", "farm", "sheep". A description of the audio data and their respective tags is provided in Section 4. The BoAW system's

http://dx.doi.org/10.21437/Interspeech.2016-839

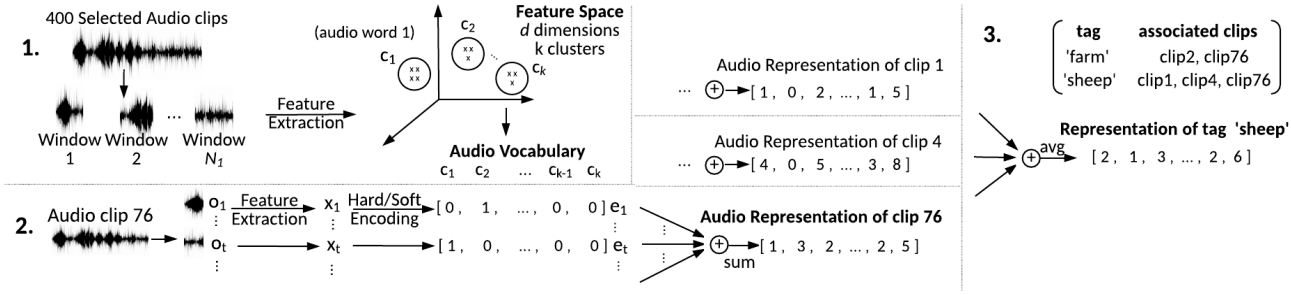

Figure 1: *System overview depicting the creation of: 1) audio-word vocabulary, 2) audio representations, and 3) tag representations.*

overview, which will be described in detail, is also presented graphically in Figure 1. First, the audio-word vocabulary is created by following the process described in Section 2.1. Then, the window encodings are computed and fused for obtaining the representation of an audio clip (see Section 2.2). Subsequently, tag representations are derived, as indicated in Section 2.3.

2.1. Audio-word vocabulary

All audio clips are converted to WAV format and resampled at 44.1 kHz. Then, a subset of the collection is randomly selected. In this work, we use a subset including 400 clips as in [1]. Each clip of the subset is partitioned in partially overlapping windows of fixed length and a feature vector is extracted from each window. The extracted features are described in Section 3. Hence, every audio 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 audio-word vocabulary.

2.2. Audio representations

In this section, we describe two schemes used for representing the semantics of audio clips with respect to the audio-word vocabulary described in Section 2.1.

2.2.1. Hard encoding

Given an audio clip, the same process for feature extraction is followed as in Section 2.1. For each window $\boldsymbol{o_t}$, a feature vector $\boldsymbol{x_t} \in R^d$ is computed (where d is the dimensionality of the feature space) and associated with the audio-word vocabulary. This is performed by assigning $\boldsymbol{x_t}$ to the closest audio word (centroid) using the Euclidean distance. The $\boldsymbol{x_t}$ vector is encoded as a k-dimensional vector $\boldsymbol{e_t}$ containing one element set to 1 and $k-1$ elements set to 0 (one-hot representation):

$$\boldsymbol{e_t} = (0, ..., 1, 0, ..., 0). \tag{1}$$

The non-zero element corresponds to the closest audio word. An entire audio clip is represented by summing the vectors computed for the respective windows. Given a collection consisting of M clips, this process results in a $M \times k$ matrix.

2.2.2. Soft encoding

Another way to calculate the encoded vector $\boldsymbol{e_t}$ is to use a soft version of the previous technique, that is more robust to noisy values. Here, we present a soft encoding scheme for formulating the $\boldsymbol{e_t}$ vector. The basic idea is to let more than one audio words to contribute to the encoding of $\boldsymbol{x_t}$. This relaxation is expected to improve the robustness of the hard encoding scheme. The contribution of the i-th audio word can be weighted via w_i ranging between 0 and 1, while the k weights sum to one. Consider the i-th audio word, i.e., the centroid of cluster $\boldsymbol{c_i}$. Assuming that the feature values corresponding to $\boldsymbol{c_i}$ follow a Gaussian distribution, we compute the respective mean $\boldsymbol{\mu_i} \in R^d$ and variance $\boldsymbol{\sigma_i^2} \in R^d$. The weight w_i is computed as:

$$w_i = \frac{p(\boldsymbol{c_i}|\boldsymbol{x_t})}{\sum_{j=1}^{k} p(\boldsymbol{c_j}|\boldsymbol{x_t})}, \tag{2}$$

and

$$p(\boldsymbol{c_j}|\boldsymbol{x_t}) = \frac{p(\boldsymbol{x_t}|\boldsymbol{c_j})p(\boldsymbol{c_j})}{p(\boldsymbol{x_t})} = \frac{p(\boldsymbol{c_j})e^{-\frac{1}{2}h_{tj}^2}}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}p(\boldsymbol{x_t})}, \tag{3}$$

where h_{tj} stands for the Mahalanobis distance between $\boldsymbol{x_t}$ and $\boldsymbol{c_j}$, $p(\boldsymbol{c_j})$ denotes the a-priori probability of cluster $\boldsymbol{c_j}$, $\boldsymbol{\Sigma}$ is the covariance matrix and $p(.)$ denotes probabilities computed via maximum likelihood estimation. We have assumed that $\boldsymbol{\Sigma}$ is diagonal, so, tying this matrix across all classes is equivalent with using the Mahalanobis distance. Hence, we derive:

$$w_i = \frac{p(\boldsymbol{c_i})e^{-h_{ti}^2}}{\sum_{j=1}^{k} p(\boldsymbol{c_j})e^{-h_{tj}^2}}. \tag{4}$$

The feature vector $\boldsymbol{x_t}$ is described by a k-dimensional vector

$$\boldsymbol{e_t'} = (w_1, w_2, ..., w_k), \tag{5}$$

where $\sum_{i=1}^{k} w_i = 1$.

2.3. Tag representations

The representation of a tag is computed by averaging the representations of the clips having this tag in their descriptions. As every audio representation is created with respect to k audio words, a tag representation will be a vector with length k. For a collection of audio clips with T (unique) tags this results in a $T \times k$ matrix. Then, Positive Pointwise Mutual Information (PPMI) weighting is applied to the matrix for obtaining more appropriate representations [16]. In addition, dimensionality reduction via Singular Value Decomposition (SVD) can be performed.

3. Fusion of feature spaces

Various feature sets have been proposed in the literature for a variety of audio signal processing applications [17], [18]. MFCCs and their temporal derivatives are the most common and extensively used features for all types of audio signals (speech, music, generic audio), however there are many others that can be exploited in order to represent a sound depending on its nature. For example features such as chroma [19], spectral flux, zero-crossing-rate, spectral centroid, etc. [20] have

shown promise for the description of musical signals. The discrimination of speech vs. non-speech has been investigated using a variety of feature representations, e.g., Linear Prediction Coefficients (LPC), Short-Time Fourier Transform STFT, modulations etc. [21], [22], [23]. Unlike MFCCs, most of these features do not work universally for all genres of audio sounds, although some intuition can be gained from the literature on the relevance of these features for speech, music and generic audio classification tasks [24]. It is necessary to include feature representations that are able to describe, discriminate and distinguish all audio genres. That is why, we experimented with three different feature spaces defined as follows:

- S_1: 13 MFCCs (concatenated with spectral energy), the 1st and 2nd order derivatives (39 features in total).
- S_2: F0 feature.
- S_3: chroma features, spectral flux, zero-crossing-rate, spectral centroid, brightness, spectral spread, spectral skewness, spectral kurtosis, roll-off (85% threshold), roll-off (95% threshold), spectral entropy, spectral flatness, roughness, irregularity, inharmonicity (27 features in total).

Since not all feature sets are equally relevant for each audio genre (e.g., S_2 is more relevant for speech, while S_3 for music), we apply weighted fusion of the features spaces, where the weights are estimated from the posterior probabilities of a 3-class audio genre classifier as described next.

Audio-word vocabulary creation: For each feature space, an audio-word vocabulary is built following the process described in Section 2.1. This process results in three vocabularies denoted as V_1, V_2, and V_3 containing k_1, k_2, and k_3 audio words corresponding to S_1, S_2, and S_3, respectively.

Genre classification: An audio clip q is categorized into one of the following classes: 1) "music", 2) "speech", and 3) "other" according to the posterior probabilities of a classifier. Here, we trained a classifier based on Support Vector Machines with linear kernel, using pyAudioAnalysis Python library [25]. The audio data for the training are presented in [26].

Encoding: The goal here is to represent an audio clip, q, in the fused S_1, S_2, and S_3. The clip q is partitioned in partially overlapping windows of a fixed size. For each window $\boldsymbol{o_t}$ a feature vector is computed with respect to $S_j, j = 1, 2, 3$. For each space S_j an encoding $\boldsymbol{e_t^j} \in R^{k_j}$ is computed according to (1). The encoded representation for each window of q is computed as the weighted concatenation of the three encodings, i.e.,

$$\boldsymbol{e_t''} = (u_1\boldsymbol{e_t^1}, u_2\boldsymbol{e_t^2}, u_3\boldsymbol{e_t^3}), \quad (6)$$

where $\sum_{i=1}^{3} u_i = 1$. The representation of an audio clip q is computed by summing the $\boldsymbol{e_t''}$ representations of the respective windows, as shown in Figure 1. The weights u_i can be set according to the classification of q to the "music", "speech" or "other" class.

4. Experimental and evaluation datasets

Experimental dataset: In total 4474 audio clips were downloaded from the online search engine Freesound [27] with the use of the Freesound API. The clips were encoded in the standard open source OGG format. These clips are not limited to only music or speech but also include sounds like footsteps, alarm notifications, street noise, etc. In general, these tend to be short clips and all of them are provided with tags and descriptions by the uploaders. Some basic statistics of the clip collection are presented in Table 1. We retained the tags that occur more than 5 times, while we discarded the tags containing only digits.

Number of clips	4474	Number of tags	37203
Min duration	0.1s	Avg tags per clip	8
Max duration	120s	Avg clips per tag	40
Avg duration	16.6s	Num of unique tags	940

Table 1: *Statistics of clip collection.*

Evaluation datasets: The task of human word semantic similarity computation was used for evaluation. In order to facilitate the comparison of the proposed approach with the related works reported in the literature of ADSMs, we used the MEN [28] and the SimLex [29] datasets. A limitation regarding the use of MEN and SimLex is the rather limited number of word pairs. In order to overcome this, we constructed two datasets including hundreds of word pairs, as presented in Table 2. The words of those pairs are included in the tag set of the clip collection described in Section 4. As groundtruth we used the similarity scores that were automatically computed via state-of-the-art CDSM and PDSM models presented in [30]. These models achieve similarity scores that are highly correlated with human ratings.

Dataset	MEN	SLex	CDSM	PDSM
# word pairs	157	44	1084	785

Table 2: *Evaluation datasets.*

5. Experiments and evaluation results

The similarity score between two words is estimated as the cosine of their respective ADSM representations. The Spearman correlation coefficient against groundtruth ratings was used as the evaluation metric. In Section 5.1, we report the evaluation results of various parameters of the baseline model. The performance of the proposed fusion according to (6) is presented in Section 5.2.

5.1. Parameters of baseline model

Here, we test the performance for each of the parameters of the BoAW model. Results (correlation) are shown in Table 3 for the MEN, SLex, CDSM, and PDSM datasets. The first line refers to the baseline proposed in [1] and the second line to the results reported in [15]. The next lines correspond to our re-implementations of the baseline model where the features of space S_1 were used. The proposed soft encoding scheme (see Section 2.2.2) was found to yield comparable performance with the hard encoding (see Section 2.2.1)[1].

The overall performance with respect to all four datasets equals to 0.365 correlation (computed by averaging the correlation scores shown in bold that yielded the best performing parameter settings).

Window length: We experimented with various values for the window length (L) used for feature extraction ranging

[1]The reported results were obtained using the hard encoding scheme. We focus on the larger datasets (i.e., CDSM, PDSM), where improvements over 0.02 to 0.03 are statistically significant, while the results for MEN, SLex are included for facilitating the comparison with the literature.

k	SVD dim	L (ms)	MEN	SLex	CDSM	PDSM
Results reported in literature [1], [15]						
100	60	250	0.402	0.233	n/a	n/a
300	-	250	0.325	0.161	n/a	n/a
Reimplementation of baseline						
100	60	250	0.382	0.302	0.321	0.294
300	-	250	0.416	0.235	0.333	0.332
100	-	25	0.397	0.327	**0.321**	0.264
		50	0.320	0.179	0.299	0.281
		100	0.373	**0.348**	0.319	0.279
		250	0.378	0.278	0.320	**0.291**
		500	**0.401**	0.286	0.307	0.280
200	-	25	0.376	**0.367**	0.356	0.320
300			**0.432**	0.355	0.365	0.311
400			0.403	0.334	0.360	0.320
500			0.398	0.285	**0.373**	**0.333**
550			0.214	0.197	0.365	0.331
300	10	25	0.300	0.329	0.364	**0.330**
	50		0.409	0.338	0.372	0.326
	90		0.435	0.332	**0.375**	0.313
	130		0.432	0.35	0.374	0.318
	170		**0.437**	**0.369**	0.371	0.315
	210		0.434	0.351	0.370	0.316

Table 3: *Correlation results of different model configurations.*

from 25 to 500ms. The window step (H) was increased (from 10 to 400ms) proportionally to the window length. Results are reported for k=100. We observe that the performance is relatively robust to a range of L values. The best correlation achieved for CDSM is 0.321, for a 25ms window, while for PDSM is 0.291 for a 250ms window[2].

Auditory dimensions: We experimented with different values for the auditory dimensions, i.e., the parameter k in k-means. The values are ranging from 100 to 550. Window length of 25ms is used with a 10ms update. As we see in Table 3, the top performance is achieved for k=500 for both CDSM and PDSM, i.e., 0.373 and 0.333, respectively.

SVD dimensions: We also experimented with the SVD dimensions regarding the dimensionality reduction of the matrix of tag representations (see Section 2.3). The model's performance is tested for SVD dimensions ranging from 10 to 210 with a step of 40. Window length of 25ms is used with a 10ms update and k=300. As we see in Table 3, dimensionality reduction slightly improves the results, i.e., from 0.365 to 0.375 for CDSM and from 0.311 to 0.313 for PDSM when reducing from k=300 to 90 dimensions.

5.2. Fusion of feature spaces

Regarding the fusion scheme described in Section 3 we used the weights presented in Table 4 according to the classification result. These weights were selected after performing an exhaustive search using held out data. The three feature spaces were computed using window length of 250ms with a 100ms update. All three of audio vocabularies are of the same size (k_1=k_2=k_3=k). In Table 5, we report the evaluation results for the fusion of the three feature spaces (S_{123}) along with the

[2]As the window length increases, the number of feature vectors (used for clustering) per audio clip decreases. So, a good practice for the building of the audio-word vocabulary would be to adjust the k parameter to the window length.

Clip categorized as	u_1	u_2	u_3
Music	0.3	0.2	0.5
Speech	0.8	0.2	0.0
Others	0.3	0.0	0.7

Table 4: *Weights for the fusion of the three feature encodings.*

Feature Space	k	SVD dim	MEN	SLex	CDSM	PDSM
S_1	300	-	0.416	0.235	0.333	0.332
S_2			0.308	0.313	0.269	0.248
S_3			0.418	0.205	0.278	0.315
S_{123}			**0.468**	**0.387**	**0.388**	**0.382**
S_1		90	0.436	0.209	0.283	0.320
S_2			0.302	0.34	0.275	0.26
S_3			0.422	0.252	0.343	0.337
S_{123}			**0.480**	**0.374**	**0.402**	**0.401**
S_1	400	-	0.457	0.24	0.298	0.309
S_2			0.304	0.334	0.283	0.259
S_3			0.423	0.300	0.384	0.343
S_{123}			**0.462**	**0.437**	**0.404**	**0.379**
S_1		90	0.427	0.317	0.375	0.331
S_2			0.314	0.351	0.278	0.254
S_3			0.46	0.225	0.293	0.302
S_{123}			**0.477**	**0.407**	**0.416**	**0.407**

Table 5: *Correlation performance of feature space fusion S_{123} vs individual encodings S_1, S_2, S_3, (L=250ms).*

performance of the individual spaces (S_1, S_2, and S_3). We observe that the proposed fusion yields higher performance than the individual spaces. For example, we achieve 12% relative improvement in correlation for MEN, 23.6% for SLex, 16.5% for CDSM, and 15.1% for PDSM (for $k = 300$, without applying SVD). Regarding the individual spaces, there is not a clear winner since S_1 and S_3 appear to achieve comparable performance (on average). The overall performance for the two large datasets (CDSM, PDSM) is 0.412 correlation (when reducing from k=400 to 90 dimensions).

6. Conclusions

In this work, we proposed the fusion of three feature spaces for constructing an ADSM and estimating the semantic distance between audio clips in the acoustic space. The fusion was based on the categorization of the content of each clip as "music", "speech" or "other". Based on the mapping between words and audio clips, this model was evaluated with respect to the computation of semantic similarity between words outperforming the baseline approach (up to 23.6% relative improvement in correlation). Also, the role of various parameters of the baseline model was investigated. It was found that the dimensionality reduction (e.g., via SVD) of the feature space can improve the performance. Regarding future work, we aim to experiment with more feature spaces and evaluate the proposed model using datasets in languages other than English. The long term goal of this work is the development of fully multimodal semantic models integrating features extracted from text, audio, and images.

7. Acknowledgements

This work has been partially supported by the BabyRobot project supported by EU H2020 (grant # 687831). The authors wish to thank Dr. Theodoros Giannakopoulos for providing the audio clip classifier.

8. References

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

[2] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet, "Audio information retrieval using semantic similarity," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[3] M. S. Lewicki, "Efficient coding of natural sounds," *Nature neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[4] S. Sundaram and S. Narayanan, "Audio retrieval by latent perceptual indexing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 49–52.

[5] M. Slaney, "Mixtures of probability experts for audio retrieval and indexing," in *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 1, 2002, pp. 345–348.

[6] L. W. Barsalou, A. Santos, K. S. W, and C. D. Wilson, "Language and simulation in conceptual processing," in *Symbols, Embodiment, and Meaning*, M. D. Vega, A. M. Glenberg, and A. C. Graesser, Eds. Oxford University Press, 2008, pp. 245–283.

[7] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, pp. 335–346, 1990.

[8] F. Vignoli and S. Pauws, "A music retrieval system based on user driven similarity and its evaluation." in *International Society for Music Information Retrieval Conference (ISMIR)*, 2005, pp. 272–279.

[9] E. Pampalk, A. Flexer, G. Widmer *et al.*, "Improvements of audio-based music similarity and genre classificaton." in *International Society for Music Information Retrieval Conference (ISMIR)*, 2005, pp. 634–637.

[10] T. Pohle, P. Knees, M. Schedl, and G. Widmer, "Automatically adapting the structure of audio similarity spaces," in *Proc. 1st Workshop on Learning the Semantics of Audio Signals (LSAS)*, 2006, pp. 66–75.

[11] J. Weston, S. Bengio, and P. Hamel, "Multi-tasking with joint semantic spaces for large-scale music annotation and retrieval," *Journal of New Music Research*, vol. 40, no. 4, pp. 337–348, 2011.

[12] E. Bruni, N. K. Tran, and M. Baroni, "Multimodal distributional semantics," *Journal of Artificial Intelligence Research*, vol. 49, no. 1, pp. 1–47, 2014.

[13] Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu, "Coherent bag-of audio words model for efficient large-scale video copy detection," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2010, pp. 89–96.

[14] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification." in *Interspeech*, 2012, pp. 2105–2108.

[15] 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.

[16] 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.

[17] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2005, pp. 34–41.

[18] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," 2004.

[19] D. P. Ellis, "Classifying music audio with timbral and chroma features," in *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, 2007, pp. 339–340.

[20] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[21] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 1997, pp. 1331–1334.

[22] B. Schuller, B. J. B. Schmitt, D. Arsić, S. Reiter, M. Lang, and G. Rigoll, "Feature selection and stacking for robust discrimination of speech, monophonic singing, and polyphonic music," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2005, pp. 840–843.

[23] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[24] Y. Lavner and D. Ruinskiy, "A decision-tree-based algorithm for speech/music classification and segmentation," *EURASIP Journal on Audio, Speech, and Music Processing*, p. 2, 2009.

[25] T. Giannakopoulos, "pyAudioAnalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[26] ——, "Study and application of acoustic information for the detection of harmful content, and fusion with visual information," *Department of Informatics and Telecommunications, vol. PhD. University of Athens, Greece*, 2009.

[27] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on multimedia*, 2013, pp. 411–412.

[28] E. Bruni, N.-K. Tran, and M. Baroni, "Multimodal distributional semantics." *J. Artif. Intell. Res. (JAIR)*, vol. 49, 2014.

[29] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, 2015.

[30] E. Iosif, S. Georgiladakis, and A. Potamianos, "Cognitively motivated distributional representations of meaning," in *10th Language Resources and Evaluation Conference (LREC)*, 2016.