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

G. Karamanolakis[1], E. Iosif[1], A. Zlatintsi[1],
A. Pikrakis[2], A. Potamianos[1]

[1]School of ECE, National Technical University of Athens, Greece
[2]Department of Informatics, University of Piraeus, Greece

Introduction

- Questions
  1. Contribution of multimodal information in lexical semantics
  2. Representation of concepts and related attributes
- Computational framework
  - Text-based Distributional Semantic Models (DSMs)
  - Bag-of-words approach
  - Semantic model based on modalities other than text?
- Audio-based DSMs (ADSMs)
  - Bag-of-audio-words approach
  - Combination of lexical features with audio clips
- Prior work
  - A. Lopopolo and E. van Miltenburg (2015)
  - D. Kiela and S. Clark (2015)

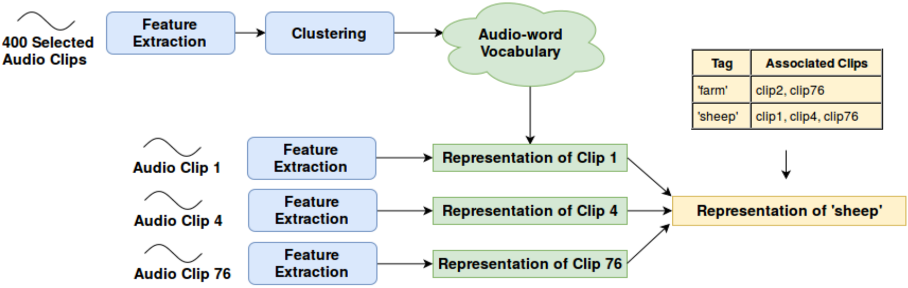
Goal – Motivation

- Goal: compute the semantic distance between words
  - Exploit their acoustic properties through the ADSM
  - Fusion of different feature encodings
- Symbol grounding problem:
  - Mainstream DSMs are ungrounded to real world
  - Rely solely on linguistic data extracted from corpora
  - Other modalities (e.g. audio,vision) contribute to the acquisition and representation of semantic knowledge
- Diversity of audio collections
  - Music, Speech, other audio classes
  - Some features do not work universally for all genres of audio sounds
  - Include feature representations that are able to describe, discriminate and distinguish all audio genres

Overview

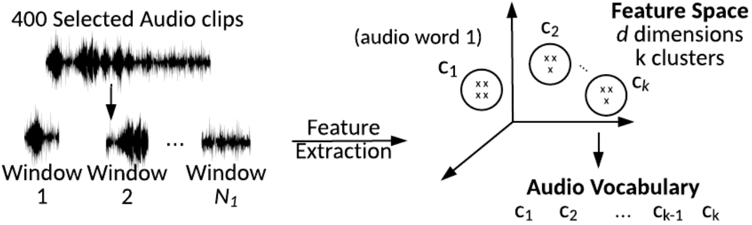
- System Description
  1. Audio-word vocabulary
  2. Audio representations
  3. Tag representations
- Fusion of feature spaces
- Experimental dataset
- Evaluation datasets
- Experiments and evaluation results
- Conclusions

Baseline System - Overview



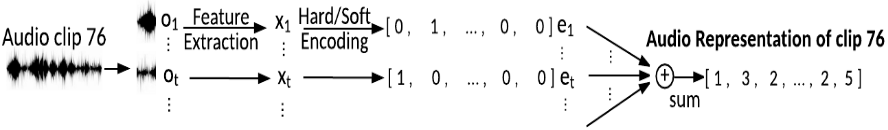
System Description - Audio-word vocabulary

- Selection of a training subset including 400 clips
- Feature extraction by partitioning clips in partially overlapping windows
- Clustering of the feature vectors (k-means)
- audio-words: the *k* centroids of the returned clusters



System Description - Audio representations (1)

- Representing the semantics of audio clips with respect to the audio-word vocabulary
- Feature extraction: For each window $\vec{o_t}$, a feature vector $\vec{x_t} \in R^d$ is computed
- Hard encoding (one-hot representation): assigning $\vec{x_t}$ to the closest audio word (centroid) using the Euclidean distance : $\vec{e_t} = (0, ..., 1, 0, ..., 0)$
- Representation of entire audio clip: summing the vectors computed for the respective windows



System Description - Audio representations (2)

- Soft encoding
  - Robust to noisy values
  - More than one audio words contribute to the encoding of $\vec{x_t}$

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

  - Weight w_i of the i_{th} audio-word:

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

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

System Description - Audio representations (3)

Soft encoding (Calculation of weights)

$$p(\vec{c}_j|\vec{x}_t) = \frac{p(\vec{x}_t|\vec{c}_j)p(\vec{c}_j)}{p(\vec{x}_t)} = \frac{p(\vec{c}_j)e^{-\frac{1}{2}h_{tj}^2}}{(2\pi)^{d/2}|\Sigma|^{1/2}p(\vec{x}_t)}, \quad (3)$$

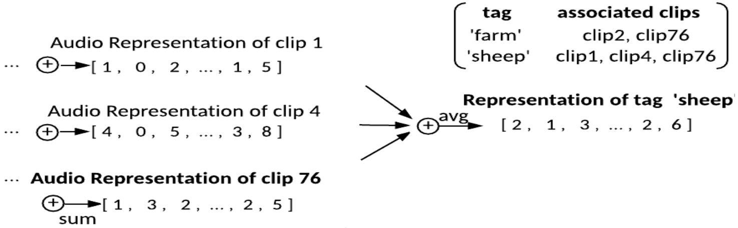
- h_{tj}: Mahalanobis distance between $\vec{x}_t$ and $\vec{c}_j$,
- $p(\vec{c}_j)$: a-priori probability of cluster $\vec{c}_j$,
- Σ: the covariance matrix,
- $p(.)$: probabilities computed via ML estimation.

By assuming Σ as diagonal:

$$w_i = \frac{p(\vec{c}_i)e^{-h_{ti}^2}}{\sum_{j=1}^{k} p(\vec{c}_j)e^{-h_{tj}^2}}. \quad (4)$$

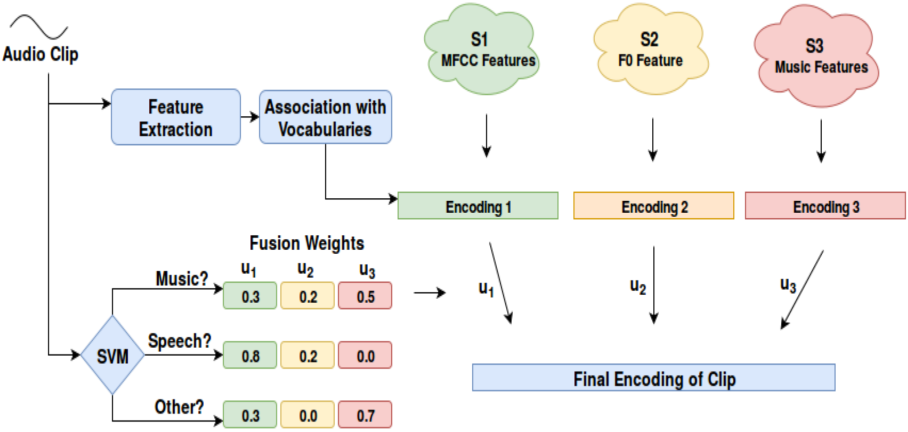
System Description - Tag representations

- Averaging the representations of the clips having this tag in their descriptions



- For a collection of clips with T (unique) tags: $T \times k$ matrix.
- Positive Pointwise Mutual Information (PPMI) weighting
- Dimensionality reduction via Singular Value Decomposition (SVD)

Fusion of feature spaces (1)



Fusion of feature spaces (2)

- Represent a sound depending on its nature
- Three different feature spaces
  - S1: 13 MFCCs, 1st and 2nd order derivatives.
  - S2: F0 feature
  - S3: chroma features, spectral flux, zero-crossing-rate, spectral centroid etc.
- Train audio-word vocabularies for each feature space
- Categorization of a clip
  - 3 classes: “music”, “speech”, “other”
  - Support Vector Machines (SVM) with linear kernel

Fusion of feature spaces (3)

- Computation of three feature encodings
  - $\vec{e_t^1}, \vec{e_t^2}, \vec{e_t^3}$ are computed with respect to S_1, S_2, S_3
- Fusion of different feature encodings
  - weighted concatenation of the three encodings:

$$\vec{e_t''} = (u_1 \vec{e_t^1}, u_2 \vec{e_t^2}, u_3 \vec{e_t^3}), \tag{5}$$

    where $\sum_{i=1}^{3} u_i = 1$.
  - Weights u_i: set according to the classification to the "music", "speech" or "other" class.
- Representation of an audio clip: summing the $\vec{e_t''}$ representations of the respective windows.

Experimental dataset

- Audio clips from the online search engine Freesound
- Not limited to only music or speech, everyday sounds e.g., footsteps, alarm notifications, street noise, etc.
- Provided with tags and descriptions by the uploaders
- Filtering of tags
  - Retain tags that occure more than 5 times
  - Discard tags that contain only digits
- Statistics of clip collection:

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

Evaluation datasets

- Evaluation task: word semantic similarity
- MEN, SimLex datasets: limited number of word pairs
- Construction of CDSM, PDSM datasets
  - State-of-the-art CDSM and PDSM models presented in [E. Iosif, S. Georgiladakis, and A. Potamianos - LREC 2016]
  - similarity scores: highly correlated with human ratings
- Statistics of evaluation datasets

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

Experimentation procedure & parameters

- Experimentation procedure
  - Similarity score between two words: cosine of their respective ADSM representations
  - Evaluation metric against ground truth ratings: Spearman correlation coefficient
- Experimentation parameters
  - L: the window length used for feature extraction (range: 25-500ms). The window step (H) increases (10-400ms) proportionally to the window length
  - k: the auditory dimensions, i.e., the k parameter of k-means (range: 100-550)
  - SVD dim: the SVD dimensions regarding dimensionality reduction of the matrix of tag representations (range: 90-300)

Evaluation results (1)

- Comparison with results reported in literature:

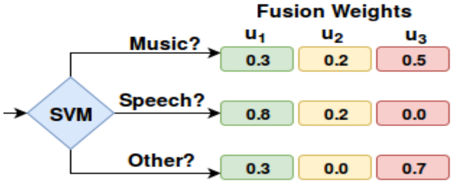
k	SVD dim	MEN	SimLex	CDSM	PDSM
Results reported in literature					
100	60	0.402	0.233	n/a	n/a
300	-	0.325	0.161	n/a	n/a
Reimplementation of baseline					
100	60	0.382	**0.302**	0.321	0.294
300	-	**0.416**	0.235	0.333	0.332

- Results reported for hard encoding (comparable performance for soft encoding)

Evaluation results (2)

- Fusion of feature spaces
  - $\vec{e}_t^{\,1}, \vec{e}_t^{\,2}, \vec{e}_t^{\,3}$ are computed with respect to S_1, S_2, S_3

- Configuring fusion weights: exhaustive search using held out data



Evaluation results (3)

- Fusion of feature spaces

Feature Space	SVD	MEN dim	SimLex	CDSM	PDSM
S_1	-	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**

Table : *Correlation performance of feature space fusion S_{123} vs individual encodings S_1, S_2, S_3, (L=250ms, $k = 300$).*

Conclusions

- Summary
  - Reimplementation of baseline ADSM described in literature
  - Investigation of various parameters of the baseline model
  - Extension of ADSM via the fusion of three feature spaces, outperforming the baseline approach (relative improvement up to 23.6%)
- Future work
  - Experiment with more feature spaces (e.g. rhythm)
  - Evaluate the proposed model using datasets in languages other than English
  - Develop fully multimodal semantic models: integration of features extracted from text, audio and images

ADSM Applications - Auto-tagging

- Comparing clips with tags?
- Bag-of-audio-words representations for both clips and tags

Clip id	Groundtruth Tags	Predicted Tags
3843	**indian**, **sitar**	**sitar**, **indian**, eastern, india, oriental
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 : Magnatagatune clips, $N = 5$ predicted tags

Thank You!