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

G. Karamanolakis\textsuperscript{1}, E. Iosif\textsuperscript{1}, A. Zlatintsi\textsuperscript{1}, A. Pikrakis\textsuperscript{2}, A. Potamianos\textsuperscript{1}

\textsuperscript{1}School of ECE, National Technical University of Athens, Greece
\textsuperscript{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

- 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

Feature Extraction → Clustering → Audio-word Vocabulary

Audio Clip 1
Audio Clip 4
Audio Clip 76

Representations of Clip 1, Clip 4, Clip 76

Tag | Associated Clips
---|-------------------
'farm' | clip2, clip76
'sheep' | clip1, clip4, clip76
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
Representing the **semantics** of audio clips with respect to the **audio-word vocabulary**

**Feature extraction**: For each window $o_t$, a feature vector $\vec{x}_t \in \mathbb{R}^d$ 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
Soft encoding
- Robust to noisy values
- More than one audio words contribute to the encoding of $\vec{x}_t$

$$\vec{e}_t = (w_1, w_2, ..., w_k), \quad (1)$$

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

$$w_i = \frac{p(\vec{c}_i|\vec{x}_t)}{\sum_{j=1}^{k} p(\vec{c}_j|\vec{x}_t)}, \quad (2)$$

where $\sum_{i=1}^{k} w_i = 1.$
System Description - Audio representations (3)

**Soft encoding** (Calculation of weights)

\[
p(\vec{c}_j | \vec{x}_t) = \frac{p(\vec{x}_t | \vec{c}_j)p(\vec{c}_j)}{p(\vec{x}_t)} = \frac{p(\vec{c}_j)e^{-\frac{1}{2}h_{ij}^2}}{(2\pi)^{d/2}|\Sigma|^{1/2}p(\vec{x}_t)},
\]

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

By assuming \(\Sigma\) as diagonal:

\[
w_i = \frac{p(\vec{c}_i)e^{-h_{ti}^2}}{\sum_{j=1}^{k}p(\vec{c}_j)e^{-h_{ij}^2}}.
\]
System Description - Tag representations

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

  \[
  \text{Audio Representation of clip 1} \\
  \quad \Rightarrow [1, 0, 2, ..., 1, 5] \\
  \text{Audio Representation of clip 4} \\
  \quad \Rightarrow [4, 0, 5, ..., 3, 8] \\
  \] \[ \text{sum} \]

  \[ \text{Audio Representation of clip 76} \\
  \quad \Rightarrow [1, 3, 2, ..., 2, 5] \]

  \[ \begin{array}{c}
  \text{tag} \\
  \quad \text{'farm'} \\
  \quad \text{'sheep'} \\
  \end{array} \]

  \[ \begin{array}{c}
  \text{associated clips} \\
  \quad \text{clip2, clip76} \\
  \quad \text{clip1, clip4, clip76} \\
  \end{array} \]

  \[ \begin{array}{c}
  \text{Representation of tag 'sheep} \\
  \quad \text{avg} \\
  \quad [2, 1, 3, ..., 2, 6] \]

- 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)

G. Karamanolakis, E. Iosif, A. Zlatintsi, A. Pikrakis, A. Potamianos

Audio-based Distributional Representations of Meaning Using a Fusion of Feature Encodings
Fusion of feature spaces (1)

Audio Clip

Feature Extraction → Association with Vocabularies

S1
MFCC Features

S2
F0 Feature

S3
Music Features

Encoding 1 → Encoding 2 → Encoding 3

Fusion Weights

Music?

Speech?

Other?

SVM

0.3 0.2 0.5

0.8 0.2 0.0

0.3 0.0 0.7

u₁

u₂

u₃

Final Encoding of Clip
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
  - $\vec{e}_1^t, \vec{e}_2^t, \vec{e}_3^t$ are computed with respect to $S_1, S_2, S_3$

- **Fusion** of different feature encodings
  - weighted concatenation of the three encodings:
    \[
    \vec{e}_t'' = (u_1\vec{e}_1^t, u_2\vec{e}_2^t, u_3\vec{e}_3^t),
    \] (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:

<table>
<thead>
<tr>
<th>Number of clips</th>
<th>4474</th>
<th>Number of tags</th>
<th>37203</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min duration</td>
<td>0.1s</td>
<td>Avg tags per clip</td>
<td>8</td>
</tr>
<tr>
<td>Max duration</td>
<td>120s</td>
<td>Avg clips per tag</td>
<td>40</td>
</tr>
<tr>
<td>Avg duration</td>
<td>16.6s</td>
<td>Num of unique tags</td>
<td>940</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MEN</th>
<th>SimLex</th>
<th>CDSM</th>
<th>PDSM</th>
</tr>
</thead>
<tbody>
<tr>
<td># word pairs</td>
<td>157</td>
<td>44</td>
<td>1084</td>
<td>785</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>$k$</th>
<th>SVD dim</th>
<th>MEN</th>
<th>SimLex</th>
<th>CDSM</th>
<th>PDSM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>60</td>
<td>0.402</td>
<td>0.233</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>300</td>
<td>-</td>
<td>0.325</td>
<td>0.161</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Results reported in literature

Reimplementation of baseline

<table>
<thead>
<tr>
<th>$k$</th>
<th>SVD dim</th>
<th>MEN</th>
<th>SimLex</th>
<th>CDSM</th>
<th>PDSM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>60</td>
<td>0.382</td>
<td>0.302</td>
<td>0.321</td>
<td>0.294</td>
</tr>
<tr>
<td>300</td>
<td>-</td>
<td>0.416</td>
<td>0.235</td>
<td>0.333</td>
<td>0.332</td>
</tr>
</tbody>
</table>

Results reported for hard encoding (comparable performance for soft encoding)
Fusion of feature spaces

\[ e_t^1, e_t^2, e_t^3 \] are computed with respect to \( S_1, S_2, S_3 \)

Configuring fusion weights: exhaustive search using held out data

---

G. Karamanolakis, E. Iosif, A. Zlatintsi, A. Pikrakis, A. Potamianos

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

<table>
<thead>
<tr>
<th>Feature Space</th>
<th>SVD</th>
<th>MEN dim</th>
<th>SimLex</th>
<th>CDSM</th>
<th>PDSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td></td>
<td>0.416</td>
<td>0.235</td>
<td>0.333</td>
<td>0.332</td>
</tr>
<tr>
<td>$S_2$</td>
<td>-</td>
<td>0.308</td>
<td>0.313</td>
<td>0.269</td>
<td>0.248</td>
</tr>
<tr>
<td>$S_3$</td>
<td></td>
<td>0.418</td>
<td>0.205</td>
<td>0.278</td>
<td>0.315</td>
</tr>
<tr>
<td>$S_{123}$</td>
<td></td>
<td>0.468</td>
<td>0.387</td>
<td>0.388</td>
<td>0.382</td>
</tr>
<tr>
<td>$S_1$</td>
<td>90</td>
<td>0.436</td>
<td>0.209</td>
<td>0.283</td>
<td>0.320</td>
</tr>
<tr>
<td>$S_2$</td>
<td></td>
<td>0.302</td>
<td>0.34</td>
<td>0.275</td>
<td>0.26</td>
</tr>
<tr>
<td>$S_3$</td>
<td></td>
<td>0.422</td>
<td>0.252</td>
<td>0.343</td>
<td>0.337</td>
</tr>
<tr>
<td>$S_{123}$</td>
<td></td>
<td>0.480</td>
<td>0.374</td>
<td>0.402</td>
<td>0.401</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Clip id</th>
<th>Groundtruth Tags</th>
<th>Predicted Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>3843</td>
<td><em>indian, sitar</em></td>
<td>sitar, <em>indian</em>, eastern, india, oriental</td>
</tr>
<tr>
<td>13526</td>
<td><em>bass, drums, drum, funky, reggae</em></td>
<td>funky, beat, <em>drums, reggae</em>, funk</td>
</tr>
<tr>
<td>15380</td>
<td><em>classical, solo, cello, violin, strings</em></td>
<td><em>cello, viola, violin, solo, classical</em></td>
</tr>
<tr>
<td>19920</td>
<td>-</td>
<td>orchestra, violins, flutes, fiddle, violin</td>
</tr>
<tr>
<td>21725</td>
<td><em>choir, choral, men, man</em></td>
<td>monks, chant, chanting, <em>men, choral</em></td>
</tr>
<tr>
<td>29231</td>
<td><em>acoustic, guitar</em></td>
<td>classical guitar, <em>guitar, acoustic</em>, lute, spanish</td>
</tr>
<tr>
<td>43390</td>
<td><em>rock, loud, pop, vocals, male vocals</em></td>
<td><em>male vocals, pop</em>, male vocal, male singer, <em>rock</em></td>
</tr>
<tr>
<td>48010</td>
<td>silence</td>
<td>low, soft, no singing, quiet, wind</td>
</tr>
<tr>
<td>57081</td>
<td><em>piano</em></td>
<td>piano solo, <em>piano</em>, classic, solo, classical</td>
</tr>
</tbody>
</table>

Table: Magnatagatune clips, $N = 5$ predicted tags
Thank You!