The Curious Case of Visual Grounding: Different Effects for Speech- and Text-based Language Encoders

UNIVERSITY OF AMSTERDAM

Adrian Sauter, Willem Zuidema, Marianne de Heer Kloots

Institute for Logic, Language and Computation, University of Amsterdam

ICASSP2026 Barcelona

Motivation

Problem

- Language models are typically unimodal
- Humans learn from **multimodal input**

Question

- Does visual grounding **improve semantic representations**?

Current Research

- Visual grounding improves semantic representations in text models [1]
- Speech models are phonetically dominated [2]

Research Gap

- No direct comparison between effects of visual grounding on speech and text models

Approach

Models

Speech-Based **L**anguage **E**ncoders (**SLE**s):

wav2vec2 [3] (ungrounded)

FaST-VGS+ [4] (visually grounded)

Text-Based **L**anguage **E**ncoders (**TLE**s):

BERT [5] (ungrounded)

VG-BERT [1] (visually grounded)

Analyses

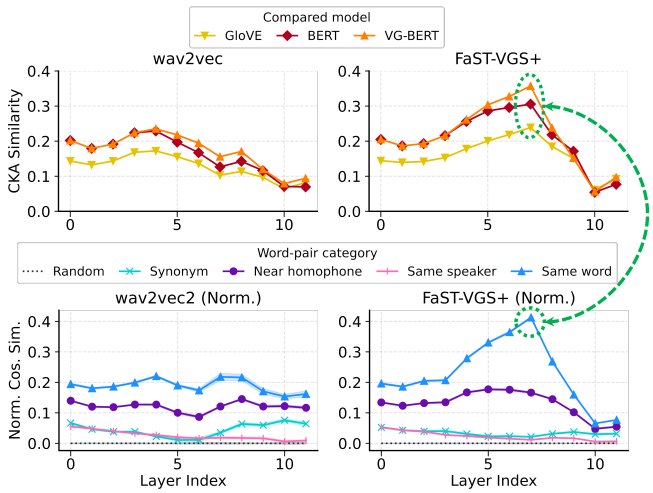
- Representation Similarity
  - Speech-text **alignment** (Centered Kernel Alignment)
- Word-Pair Similarity
  - Probes **identity vs. meaning vs. phonetics**
  - Cosine similarity between pairs of same word/ synonym/ homophones/ same speaker
- Clustering
  - **New datasets**: controlled groups of **semantically similar and phonetically distinct words**, or vice versa
    - e.g., [piano, guitar, violin], [apple, banana, orange]
    - e.g., [liquor, kicker, ticker], [chin, chip, chuck]
  - Measure **clustering quality** with silhouette coefficient
- Subspace Probing
  - Information **decodability** (Linear Discriminant Analysis)

Takeaways & Future Work

- Visual grounding aligns speech and text representations — but improves word identity, not meaning, in speech models
- Visual grounding benefits semantic structure in text models, but disrupts it in speech models
- Can visual grounding improve speech semantics by targeting semantically relevant subspaces (e.g., middle layers)?

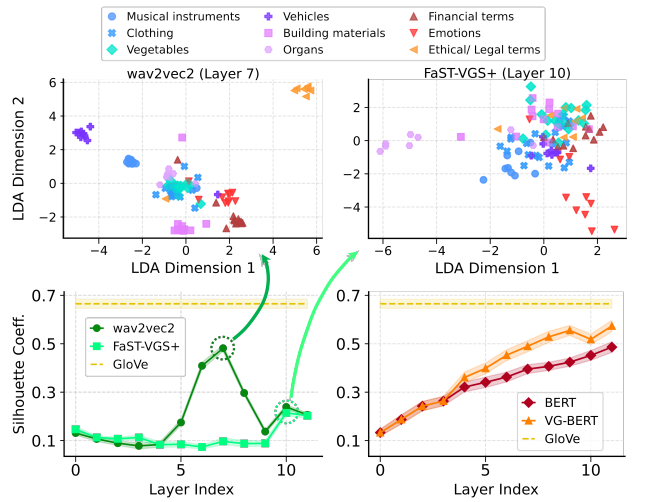
Findings

Global representational comparisons & word pairs



Finding 1: Visual grounding bridges modalities by enhancing word identity representation

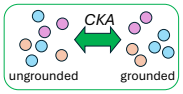
Semantic clustering



Finding 2: Visual grounding does not improve semantic clustering in SLEs, even though it does so in TLEs

Additional findings

- **Different impact** on representation geometry
  - TLEs: grounded ≈ ungrounded (CKA ~0.75)
  - SLEs: more restructuring (CKA ~0.46)



- Link between change and semantics (corr. sil. coeff. and CKA):
  - Neg. correlation in SLEs: **more change → worse clustering**
  - Pos. correlation in TLEs: **more change → better clustering**
- Concrete word groups (e.g., vegetables) **cluster better** than abstract ones (e.g., emotions)
  - Effect is even stronger with visually grounding

[1] Yizhen Zhang, Minkyu Choi, Kuan Han, and Zhongming Liu, "Explainable semantic space by grounding language to vision with cross-modal contrastive learning," NeurIPS 2021, vol. 34, pp. 18513–18526.
[2] Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe, "Self-Supervised Speech Representations are More Phonetic than Semantic," in Interspeech 2024, pp. 4578–4582.
[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," NeurIPS 2020, vol. 33, pp. 12449–12460.
[4] Puyuan Peng and David Harwath, "Self-supervised Representation Learning for Speech Using Visual Grounding and Masked Language Modeling," in Proceedings of the AAAI Symposium on AI for Speech and Audio Processing (AAAI-SAS), 2022.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in NAACL 2019. pp. 4171–4186, Association for Computational Linguistics.