Multimodal Compositional Distributional Semantics

Saba Nazir

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science
University College London

August 11, 2025

I, Saba Nazir, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Multimodal Compositional Distributional Semantics
Saba Nazir

Representing meaning in language has long been a key challenge in natural language processing, with diverse approaches seeking to capture its complexity. Distributional semantics offers a methodology for training high quality statistical representations for words; compositional distributional semantics extends these to longer phrases and sentences by encoding the statistics of words with function types, such as adjectives and verbs. Multimodal distributional semantics combines linguistic statistics with visual and auditory perceptions to ground word representations. While successful in word-level tasks, particularly in visual contexts, its application to compositional semantics with auditory grounding remains largely unexplored. This thesis addresses this limitation by introducing a multimodal compositional distributional semantics framework that builds upon tensor-based compositional models and grounds them auditorily. To the best of our knowledge, this is the first work of its kind. The framework is evaluated using a newly developed sound-relevant adjective-noun phrase similarity benchmark, measuring semantic and audio similarity. Results show that (1) compositional models outperform non-compositional baselines, (2) matrix-based compositions surpass vector addition and multiplication, and (3) multimodal models enhance performance over unimodal ones. Further evaluations on a multi-label sentiment classification task demonstrates improved accuracy over text-only models. Additionally, this thesis provides a general baseline for the application of multimodal distributional semantics in recommendation systems, while opening new avenues for future research.

Impact Statement

This thesis presents a novel multimodal language composition approach, combining audio and text to improve language representations. It has broad applications across various fields:

- Within academia, to our best knowledge, this is the first work to integrate audio-textual cues into language composition, advancing distributional compositional semantics by using multimodal data with type-driven approaches. It sets a strong foundation for future research in multimodal learning. Additionally, it introduces a novel multimodal phrase similarity benchmark that captures both semantic and audio similarities between phrases.

- Beyond academia, it can enhance systems in industries such as media, entertainment, and e-commerce. For instance, in audio captioning systems, to generate more accurate descriptions of audio content, improving accessibility and user experiences on media platforms. Similarly, in recommendation systems, to allow for more personalised suggestions in services like music streaming and video platforms.

- The societal impact of this research is equally significant. For instance, by improving sentiment analysis, this work can benefit industries such as marketing, and social media monitoring, enabling more effective understanding of public sentiment expressed through multimedia.

In summary, this thesis makes both theoretical and practical contributions to the field of NLP, with far-reaching implications for academia, industry, and society.

Acknowledgements

Completing this PhD has been a wild ride, and I am deeply thankful to everyone who stood by me along the way. Firstly, a huge thanks to my supervisor, *Mehrnoosh Sadrzadeh*, for her guidance, expertise, and constant support. I could not have asked for a more dedicated mentor, and I will always be grateful for your contributions to my growth as a researcher. I also extend my sincere thanks to *Stephen Clark* for his insightful suggestions and comments, which greatly enriched my work. To my colleagues and fellow researchers, thank you for stimulating discussions, honest feedbacks, and all the laughs that we shared throughout this journey.

A special thanks to my parents for their constant love, prayers, and sacrifices. To my husband, *Rehan*, you've been my rock through this rollercoaster, keeping me grounded with your love, patience, and endless silly jokes. To my superhero, *Haadi*, thank you for making me smile even on the hardest days and reminding me to celebrate the small joys in life. To my extended family, friends, and to those who believed in me even when I doubted myself—this achievement is just as much yours as it is mine.

This research was mainly supported by UCL Research Studentship and Department of Computer Science, University College London.

UCL Research Paper Declaration Form: Referencing the Doctoral Candidate's Own Published Work(s)

1. **For a research manuscript that has already been published (if not yet published, please skip to section 2):**

   (a) **What is the title of the manuscript?** The Potential of Multimodal Compositionality for Enhanced Recommendations through Sentiment Analysis

   (b) **Please include a link to or doi for the work:** https://doi.org/10.1145/3686215.3690145

   (c) **Where was the work published?** Companion Proceedings of the 26th ACM International Conference on Multimodal Interaction (ICMI)

   (d) **Who published the work?** ACM

   (e) **When was the work published?** November 2024

   (f) **List the manuscript's authors in the order they appear on the publication:** Saba Nazir, Mehrnoosh Sadrzadeh

   (g) **Was the work peer-reviewed?** Yes

   (h) **Have you retained the copyright?** Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g., medRxiv)? If 'Yes,' please give a link or doi:** No

2. **For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3):**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server (e.g., medRxiv)? If 'Yes,' please give a link or doi:**

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4):**

   Saba Nazir: Framework development; Data collection, preparation, and analysis; Execution of computations and result interpretation.

   Mehrnoosh Sadrzadeh: Overall supervision; Expert guidance in computational linguistics; Critical feedback and refinement of methodologies.

4. **In which chapter(s) of your thesis can this material be found?**
   Chapter 3 (Methods), Chapter 5 (Evaluation) and Chapter 6 (Sentiment Analysis).

e-Signatures Confirming That the Information Above Is Accurate:

Candidate: Saba Nazir **Date:** February 2025

Senior Author: Mehrnoosh Sadrzadeh **Date:** February 2025

UCL Research Paper Declaration Form: Referencing the Doctoral Candidate's Own Published Work(s)

1. **For a research manuscript that has already been published (if not yet published, please skip to section 2):**

   (a) **What is the title of the manuscript?** How Does an Adjective Sound Like? Exploring Audio Phrase Composition with Textual Embeddings

   (b) **Please include a link to or doi for the work:** `https://aclanthology.org/2024.clasp-1.3/`

   (c) **Where was the work published?** Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning

   (d) **Who published the work?** Association for Computational Linguistics (ACL)

   (e) **When was the work published?** October 2024

   (f) **List the manuscript's authors in the order they appear on the publication:** Saba Nazir, Mehrnoosh Sadrzadeh

   (g) **Was the work peer-reviewed?** Yes

   (h) **Have you retained the copyright?** Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g., medRxiv)? If 'Yes,' please give a link or doi:** No

2. **For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3):**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server (e.g., medRxiv)? If 'Yes,' please give a link or doi:**

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4):**

   Saba Nazir: Framework development; Data collection, preparation, and analysis; Execution of computations and result interpretation.

   Mehrnoosh Sadrzadeh: Overall supervision; Expert guidance in computational linguistics; Critical feedback and refinement of methodologies.

4. **In which chapter(s) of your thesis can this material be found?**
   Chapter 3 (Methods) and Chapter 5 (Evaluation).

e-Signatures Confirming That the Information Above Is Accurate:

Candidate: Saba Nazir **Date:** February 2025

Senior Author: Mehrnoosh Sadrzadeh **Date:** February 2025

UCL Research Paper Declaration Form: Referencing the Doctoral Candidate's Own Published Work(s)

1. **For a research manuscript that has already been published (if not yet published, please skip to section 2):**

   (a) **What is the title of the manuscript?** Semantic and Lexical Token Based Vectors Improve Precision of Recommendations for TV Programmes

   (b) **Please include a link to or doi for the work:** `https://doi.org/10.xxxx/xxxxxx`

   (c) **Where was the work published?** IEEE International Symposium on Multimedia (ISM)

   (d) **Who published the work?** IEEE

   (e) **When was the work published?** December 2023

   (f) **List the manuscript's authors in the order they appear on the publication:** Taner Cagali, Hadi Wazni, Saba Nazir, Mehrnoosh Sadrzadeh, Chris Newell

   (g) **Was the work peer-reviewed?** Yes

   (h) **Have you retained the copyright?** No

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g., medRxiv)? If 'Yes,' please give a link or doi:** No

2. **For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3):**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server (e.g., medRxiv)? If 'Yes,' please give a link or doi:**

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

(e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4):**

   Taner Cagali & Hadi Wazni: Data collection, preparation, and analysis for monthly dataset; Framework development and execution.

   Saba Nazir: Responsible for weekly dataset processing and implementation.

   Mehrnoosh Sadrzadeh & Chris Newell: Overall supervision and expert guidance in computational linguistics and recommendation; Critical feedback and refinement of methodologies.

4. **In which chapter(s) of your thesis can this material be found?**
   Chapter 7 (Multimodal Recommendations).

e-Signatures Confirming That the Information Above Is Accurate:

Candidate: Saba Nazir **Date:** February 2025

Senior Author: Mehrnoosh Sadrzadeh **Date:** February 2025

UCL Research Paper Declaration Form: Referencing the Doctoral Candidate's Own Published Work(s)

1. **For a research manuscript that has already been published (if not yet published, please skip to section 2):**

   (a) **What is the title of the manuscript?** Audiovisual, Genre, Neural and Topical Textual Embeddings for TV Programme Content Representation

   (b) **Please include a link to or doi for the work:** `https://ieeexplore.ieee.org/document/9327970`

   (c) **Where was the work published?** IEEE International Symposium on Multimedia (ISM)

   (d) **Who published the work?** IEEE

   (e) **When was the work published?** December 2020

   (f) **List the manuscript's authors in the order they appear on the publication:** Saba Nazir, Taner Cagali, Mehrnoosh Sadrzadeh, Chris Newell

   (g) **Was the work peer-reviewed?** Yes

   (h) **Have you retained the copyright?** No

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g., medRxiv)? If 'Yes,' please give a link or doi:** No

2. **For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3):**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server (e.g., medRxiv)? If 'Yes,' please give a link or doi:**

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4):**

   Saba Nazir: Data collection, preparation, and analysis for audio/visual dataset dataset; Framework development and execution.

   Taner Cagali: Responsible for subtitle vectorization and processing.

   Mehrnoosh Sadrzadeh & Chris Newell: Overall supervision and expert guidance in computational linguistics and recommendation; Critical feedback and refinement of methodologies.

4. **In which chapter(s) of your thesis can this material be found?**
   Chapter 7 (Multimodal Recommendations).

e-Signatures Confirming That the Information Above Is Accurate:

Candidate: Saba Nazir **Date:** February 2025

Senior Author: Mehrnoosh Sadrzadeh **Date:** February 2025

UCL Research Paper Declaration Form: Referencing the Doctoral Candidate's Own Published Work(s)

1. **For a research manuscript that has already been published (if not yet published, please skip to section 2):**

   (a) **What is the title of the manuscript?** Cosine Similarity of Multimodal Content Vectors for TV Programmes

   (b) **Please include a link to or doi for the work:** `https://arxiv.org/pdf/2009.11129`

   (c) **Where was the work published?** Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR.

   (d) **Who published the work?** PMLR (Proceedings of Machine Learning Research)

   (e) **When was the work published?** 2020

   (f) **List the manuscript's authors in the order they appear on the publication:** Saba Nazir, Taner Cagali, Chris Newell, Mehrnoosh Sadrzadeh

   (g) **Was the work peer-reviewed?** Yes

   (h) **Have you retained the copyright?** No

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g., medRxiv)? If 'Yes,' please give a link or doi:** No

2. **For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3):**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server (e.g., medRxiv)? If 'Yes,' please give a link or doi:**

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4):**

   Saba Nazir: Data collection, preparation, and analysis for audio/visual dataset dataset; Framework development and execution.

   Taner Cagali: Responsible for subtitle vectorization and processing.

   Mehrnoosh Sadrzadeh & Chris Newell: Overall supervision and expert guidance in computational linguistics and recommendation; Critical feedback and refinement of methodologies.

4. **In which chapter(s) of your thesis can this material be found?**
   Chapter 7 (Multimodal Recommendations).

e-Signatures Confirming That the Information Above Is Accurate:

Candidate: Saba Nazir **Date:** February 2025

Senior Author: Mehrnoosh Sadrzadeh **Date:** February 2025

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *Language is not just a body of vocabulary or a set of grammatical rules. Every language is an old-growth forest of the mind.*
>
> — **Wade Davis**

Language is more than a mere sequence of words; it is a medium rich with context and perception. On the contextual side, traditional **distributional semantics** (Harris [1], and Firth [2]), argues that meanings of words can be deduced from the contexts in which they frequently occur. This argument has led to the development of methodologies for learning high-quality vector representations for words [3, 4]. Over time, these models evolved, from simple word co-occurrence matrices (e.g., LSA [5]), to advanced neural embeddings (such as Skip-gram [6], GloVe [7], and BERT [8]). While these advancements have led to significant achievements in word-level representation tasks, such as word similarity and relatedness [9–11], they struggle to capture complex linguistic structures like phrases and sentences.

To overcome the limitations of distributional semantics, **compositional distributional semantics** extends its principles by incorporating the *compositionality principle* [12]. This principle suggests that the meaning of complex expressions arises from the meanings of their components and the rules governing their combination. Early compositional distributional approaches employed straightforward operations like vector addition and pointwise multiplication [13, 14], but these techniques were limited due to their commutative nature. As the field advanced, more sophisticated methods emerged, including neural network-based models for sentence representa-

tions [15–17] and tensor-based compositional models [18, 19], offering improved flexibility and expressiveness in representing linguistic structures. A significant advancement is the categorical compositional framework by Coecke et al. [20], which elegantly unified category theory with distributional semantics, laying a robust mathematical foundation for modelling compositional meaning. Other inspiring works in this context include those by Baroni & Zamparelli [21], Maillard & Clark [22], Grefenstette et al. [23], and Wijnholds et al. [24]. These methods have been evaluated using both traditional co-occurrence representations and neural embeddings, often outperforming simpler operators and non-compositional baselines [24, 25].

One of the main criticisms of distributional semantics is its lack of grounding in real-world knowledge, such as *perceptual* data from auditory and visual experiences. Purely textual models lack grounding in sensory modalities, and hence fall short of human-like semantic understanding [26]. For example, they may struggle to distinguish between *loud explosion* and *bright explosion*, lacking access to auditory or visual cues. This has led to **multimodal distributional semantics**, which integrates sensory data with text. Building on the work of Feng & Lapata [27], later studies [28–30] showed that grounding text in visual features improves semantic representations.

Kiela & Clark [31] and Lopopolo & Miltenburg [32] extended this by integrating audios instead, enriching semantic grounding for sound-related concepts like *rain* and *guitar*. Later, Kiela & Clark [33] used deep learning to surpass bag-of-audio-words (BoAW), achieving superior performance on the audio variant of MEN dataset [11], inspiring other neural architectures, document-level tasks, and multimodal applications [34–36].

Despite these advancements, extending grounded distributional semantics to compositional models remains an underexplored area. Recent studies, including Lewis et al. [37] and Wazni et al. [38], focused on incorporating images into compositional frameworks, demonstrating the potential of grounded compositions to surpass state-of-the-art models like CLIP [39]. However, no such efforts have been made to integrate audio into compositional distributional semantics. Figure 1.1 shows the transformation of distributional paradigms over time.

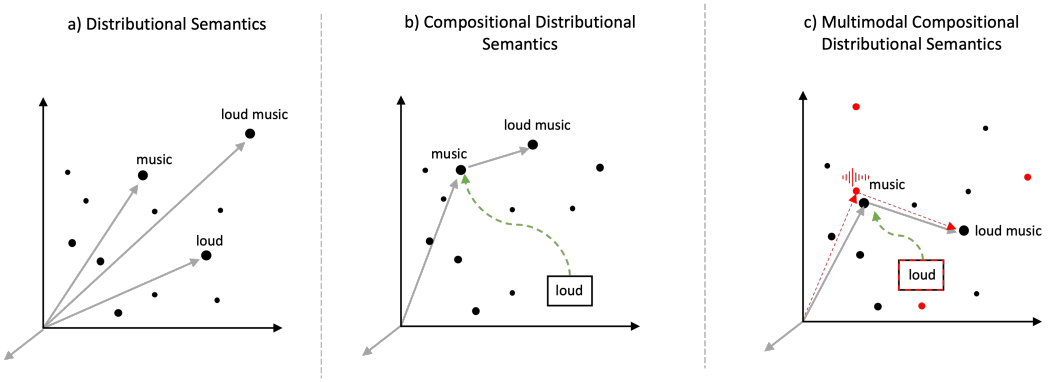


Figure 1.1: A hypothetical vector space illustration of the concepts from distributional semantics to multimodal compositional distributional semantics.

To addresses this limitation, this thesis proposes a formalism called **MultiCoDi**, a multimodal compositional distributional semantics model, with an aim to ground the existing compositional frameworks, such as those by Coecke et al. [20] and Maillard et al. [40], with auditory data. While Kiela et al. [33] laid the foundation for bimodal audio-text learning at the word level, their approach has yet to be extended to complex linguistic structures like phrases and sentences. The proposed framework addresses it by integrating purely textual compositional frameworks with multimodal information. For example, consider adjective-noun phrases, Maillard and Clark [22] introduced a tensor-based compositional framework where nouns are represented as vectors and adjectives as matrices. MultiCoDi extends it by grounding these representations in both textual and auditory modalities. Specifically, nouns are *grounded vectors* learnt jointly from the statistics of their occurrences in text and the auditory information encoded in the sound files associated with them, but more importantly, that adjectives are *grounded matrices* also jointly learnt from text statistics and audio data. For example, for the following composition:

$$\overrightarrow{loud\ music} = \overline{loud} \times \overrightarrow{music}$$

$\overline{loud}$ is the matrix representation of the adjective *loud*, $\overrightarrow{music}$ is the vector representation of the noun *music*, and $\times$ denotes matrix-vector multiplication, yielding a phrase embedding that integrates textual and auditory data.

One of the main reasons that audio-textual learning remains an underserved

area is the scarcity of resources. For example, existing word similarity benchmarks, such as MEN [29], WordSim353 [10], and SimLex [9], are predominantly textual, focusing on frequently occurring words in language which often lack sensory or auditory relevance. Similarly, phrase datasets, such as Mitchell & Lapata [41] and Vecchi [42], prioritise generic English adjectives, such as *red, new* and *early*, offering limited coverage of sound-related concepts like *melody* and *creaky*. While multimodal datasets like AudioCaps [43] and Clotho [44] link audio to sentence-level captions, they lack the granularity needed for phrase-level analysis.

This thesis addresses the above gap by introducing a sound-relevant textual-audio phrase similarity dataset. Since the concepts in this dataset are tied to auditory relevance, two distinct measures of similarity are proposed: (1) semantic similarity and (2) auditory similarity, and their proposed datasets are called **SemPhrase** and **AudPhrase** respectively. For example, in the phrase pair *creaky door* and *creaky bridge*, the phrases may have low semantic similarity but high auditory similarity. The dataset ensures a clear distinction between these two dimensions to better understand their relationship. Using this dataset, a comparative study is conducted by grounding the composition methods proposed by Mitchell & Lapata [41], Baroni & Zamparelli [21], and Maillard & Clark [22]. The results show that (1) compositional frameworks outperform non-compositional approaches, (2) matrix-based composition methods outperform vector-based approaches, and (3) most importantly, multimodal models achieve better performance than unimodal models.

What is Sound Relevance? Sound-relevant words are those that evoke or are inherently associated with specific sounds, such as *creak*, *crunch*, or *roar*. Within this category, sound-relevant adjectives are words that, when combined with nouns, describe or emphasise auditory characteristics, as in phrases like *creaky door* or *loud horn*.

Beyond similarities, this thesis extends the proposed **MultiCoDi** framework to sentiment analysis. Traditional models often focus on individual words, overlooking how sentiment emerges from phrase and sentence composition. For instance, in *not happy*, the word *happy* conveys positive sentiment, but *not* reverses it. Moilanen et

al. [45] incorporated grammatical rules into sentiment analysis rather than relying solely on word counts. Yessenalina et al. [46] introduced phrase-level sentiment analysis using matrix-space models inspired by Baroni and Zamparelli [21], learning matrices for all words. Asaadi et al. [47] further refined these models by optimizing matrices using unigram and bigram patterns. While these advancements improved textual sentiment analysis, recent research has shifted towards **multimodal sentiment analysis**, which integrates multiple modalities. Chen et al. [48] developed an image sentiment classifier using adjective-noun pairs from image tags, while Li et al. [49] translated images into textual descriptions for sentiment prediction. However, the integration of audio in sentiment analysis remains largely unexplored. This thesis addresses this gap by applying MultiCoDi to a multi-label sentiment classification task, integrating audio and textual embeddings into a neural network for sentiment prediction. Experimental results show that multimodal compositional approaches yield stronger correlations with human judgments than unimodal models.

Finally, this thesis takes the first steps to apply the proposed framework to the media recommendations. Modern recommender systems utilise vector semantics, representing words and documents as high-dimensional vectors. However, many still rely on single-modal data, fail to fully exploit the potential of multimodal integration. For instance, Yang et al. [50] relied solely on tags and titles, while Ekenel et al. [51] combined images with tags, offering limited integration of diverse modalities. Bougiatiotis and Giannakopoulos [52] attempted to incorporate audio, video, and subtitles, but their approach overlooked genre information, a critical factor for context-aware recommendations. This thesis aims to address these shortcomings by taking inspiration from the multimodal approach proposed by Kiela & Clark [33]. It extends it from word-level to document-level, further enriched with genre and visual vectors for **multimodal recommendations** of TV programmes. Experiments demonstrated that the inclusion of multimodal information consistently outperformed single-modal approaches, achieving more precise and diverse recommendations.

Thesis Structure

The rest of the thesis is structured as follows:

- Chapter 2 reviews key literature, covering distributional semantics, compositionality, multimodal grounding, auditory feature extraction, and evaluation benchmarks for unimodal and multimodal compositional semantics.
- Chapter 3 presents the core contribution of this thesis: the MultiCoDi framework, a type-driven compositional model integrating auditory and textual data. It extends existing models such as linear regression and tensor skip-gram to incorporate auditory features. This chapter is partly published in Nazir & Sadrzadeh [53].
- Chapter 4 introduces a novel multimodal dataset for evaluating phrase-level semantic and auditory similarities, addressing key gaps in existing benchmarks. It details the methodology for dataset construction and human annotations. Information about this dataset appeared in Nazir & Sadrzadeh [53] and [54].
- Chapter 5 assesses the quantitative and qualitative effectiveness of the proposed compositional models in capturing both semantic and auditory relationships between phrases. The integration of auditory information with textual data is shown to significantly enhance model performance compared to unimodal approaches. These results are published in Nazir & Sadrzadeh [53].
- Chapter 6 extends the application of MultiCoDi in sentiment analysis by leveraging textual and auditory data to address critical limitations in traditional approaches. This chapter is partly published in Nazir & Sadrzadeh [54].
- Chapter 7 explores the integration of multimodal distributional semantics in TV programme recommendations. This chapter is partly published in Nazir et al. [55, 56] and Cagali et al. [57]
- Finally, Chapter 8 summarises the key contributions, highlighting advancements in compositional distributional semantics and the broader impact of integrating auditory cues with textual data. It also outlines promising directions for future research.

Published Contributions

The following publications have resulted from this research:

Saba Nazir and Mehrnoosh Sadrzadeh. “The Potential of Multimodal Compositionality for Enhanced Recommendations through Sentiment Analysis”. In: *Companion Proceedings of the 26th ACM ICMI International Conference on Multimodal Interaction (ICMI)*. 2024.

Saba Nazir and Mehrnoosh Sadrzadeh. “How Does an Adjective Sound Like? Exploring Audio Phrase Composition with Textual Embeddings”. In: *Proceedings of the 2024 Conference on Multimodality and Interaction in Language Learning (CLASP), ACL Anthology.* 2024.

Taner Cagali, Hadi Wazni, Saba Nazir, Mehrnoosh Sadrzadeh, and Chris Newell. “Semantic and Lexical Token Based Vectors Improve Precision of Recommendations for TV Programmes”. In: *IEEE International Symposium on Multimedia (ISM)*. 2023.

Saba Nazir, Taner Cagali, Chris Newell, and Mehrnoosh Sadrzadeh. “Audiovisual, genre, neural and topical textual embeddings for TV programme content representation”. In: *IEEE International Symposium on Multimedia (ISM)*. 2020.

Saba Nazir, Taner Cagali, Chris Newell, and Mehrnoosh Sadrzadeh. “Cosine Similarity of Multimodal Content Vectors for TV Programmes”. In: *Proceedings of the 37th International Conference on Machine Learning (ICML) Media Discovery (ML4MD) Workshop*. PMLR 108. 2020.

Chapter 2

Background

This chapter provides a comprehensive overview of the literature on multimodal compositional distributional semantics. It begins with an exploration of distributional semantics, followed by a survey on compositional distributional semantics. The chapter further examines the state-of-the-art embedding techniques for audio learning, reviews multimodal compositional distributional approaches, and highlights fusion techniques as well as the evaluation datasets. Finally, it concludes with a discussion of the DisCoCat and its extension to CCG.

2.1 Distributional Semantics

Distributional semantics is founded on the principle that the meaning of a word can be inferred from the contexts in which it frequently occurs. This foundational concept was introduced by Harris [1] and further popularised by Firth [2], who stated that "a word is characterised by the company it keeps". More precisely, the **distributional hypothesis**, posits that "words that occur in similar contexts tend to have similar meanings". This principle has become the foundation for vector space models, which mathematically capture word meanings by representing them as multidimensional vectors based on their contexts. Over time, these models have undergone significant advancements, as detailed below:

2.1.1 Classical Distributions

The foundation of distributional semantics lies in *count-based models*, which infer word meaning from co-occurrence patterns within a fixed context window. For a

target word w_i, surrounding words $w_{i-n}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+n}$ are used to construct a co-occurrence matrix, where each entry records how often a context word appears near w_i. Function words (e.g., "the", "and") are typically excluded to focus on content-bearing terms. Formally, word meaning is represented as:

$$\overrightarrow{\text{word}} = \sum_i c_i \vec{n_i},$$

where c_i is the co-occurrence strength and $\vec{n}_i$ is the one-hot (basis) vector for context word n_i. While this may seem circular, early models treat context vectors as fixed, and meaning emerges from comparing co-occurrence distributions rather than assuming predefined semantics. For example, *dog* may co-occur with *bark*, *tail*, and *pet*, and be approximated as:

$$\overrightarrow{\text{dog}} = c_1 \vec{\text{bark}} + c_2 \vec{\text{tail}} + c_3 \vec{\text{pet}}.$$

Methods like Latent Semantic Analysis (LSA) [5] and Hyperspace Analogue to Language (HAL) [58] apply matrix decomposition to extract latent structure. Weighting schemes such as Pointwise Mutual Information (PMI) and PPMI further refine raw counts by highlighting informative associations and reducing noise [59].

2.1.2 Neural Embeddings

Count-based models struggled with sparsity and generalization, leading to neural embeddings that generate dense representations and capture complex relationships through contextual adjustments. Key models are discussed below:

1. *Skip-gram* [6]: The Skip-gram method, commonly implemented as part of *word2vec*[1], is designed to maximise the likelihood of predicting context words given a target word. For example, in a sentence like *the bird sang in the tree*, if the target word is *bird*, the model learns to predict context words such as *the*, *sang*, *in*, and *the* within a defined context window around the target word.

   Skip-gram with Negative Sampling (SGNS) [60] is introduced to address the computational inefficiency of training on all possible context words. Instead of updating weights for all words in the vocabulary, SGNS randomly samples a

[1] github.com/tmikolov/word2vec

small set of negative words (words that are not part of the actual context) and updates the weights to distinguish between the actual context and the negative samples. This significantly reduces the computational cost, making it much more efficient with the following objective function:

$$\sum_{c \in C} \log \sigma(n \cdot c) + \sum_{\bar{c} \in \bar{C}} \log \sigma(-n \cdot \bar{c})$$

Where C is the set of context words associated with the target word n, while $\bar{C}$ is the set of negative samples drawn from a unigram distribution raised to a certain power (typically 3/4). The function σ stands for the sigmoid function, which estimates the probabilities that context words appear given the target word.

2. *GloVe* [7]: The primary idea behind GloVe (Global Vectors for Word Representation) is to utilise the co-occurrence matrix of words within a large corpus, which counts how often words appear together in a given context. This model constructs word embeddings by factorizing this matrix into lower-dimensional vectors that capture semantic relationships between words. Unlike Word2Vec, which focuses on local context, GloVe leverages both local and global co-occurrence statistics.

3. *FastText* [61]: FastText enhances word representations by incorporating subword information. It employs a methodology similar to the Skip-gram model but enriches its predictions by utilising character n-grams derived from the target word. For instance, the word *apple* is decomposed into n-grams such as *app, ap, pl, and le*. The word's embedding is computed by averaging the vectors of these n-grams along with a distinct vector for the word itself. This allows FastText to generate embeddings for out-of-vocabulary words by leveraging the embeddings of their constituent n-grams.

4. *ELMo* [62]: ELMo utilises a bidirectional Long Short-Term Memory (bi-LSTM) network trained on a language modelling task to generate dynamic, context-sensitive word embeddings. This model allows word representations to adapt based on their contextual usage within sentences. Final embeddings are derived from the hidden states of the bi-LSTM, effectively integrating information from

both forward and backward contexts. Each word's embedding is a weighted sum of these hidden states, capturing a rich spectrum of contextual influences.

5. *BERT* [8]: Utilising the transformer architecture, BERT employs a bidirectional training approach that simultaneously considers the left and right context of a word within a sentence, allowing it to capture nuanced meanings and relationships that unidirectional models cannot. It is pre-trained on vast corpora using two main tasks: the Masked Language Model (MLM), where random words in a sentence are masked for prediction, and Next Sentence Prediction (NSP), which helps the model understand sentence relationships. The adaptability of BERT for fine-tuning on specific downstream tasks has made it a foundational model in NLP, influencing subsequent research and applications in understanding language semantics and context.

2.2 Compositional Distributional Semantics

Traditional symbolic approaches in formal semantics, such as those introduced by Montague [12] and further expanded by Dowty et al. [63], address the principle of compositionality by pairing syntactic structures with semantic interpretation rules. These methods rely on formal grammatical frameworks, such as categorial grammars, to systematically compute the meaning of complex linguistic expressions. While symbolic approaches excel at producing logical and interpretable representations of meaning, they often struggle to account for the subtleties of meaning derived from context and usage. For instance, the phrase *barked at* conveys a different action than *growled at*, even though both phrases share similar syntactic structures.

In contrast, distributional semantics represents word meanings as vectors, positioning words with similar contexts closer together in a shared semantic space to capture their contextual similarity. However, while distributional semantics is effective for word-level representations, such as identifying the similarity between *dog* and *puppy*, it lacks a robust mechanism for combining these representations into meaningful representations of phrases or sentences. For example, while it can model similarity between individual words, it fails to capture how the meaning of

happy dog differs from just combining *happy* and *dog*, or how word order and syntax contribute to the meaning of a sentence like *The dog barked*.

Compositional Distributional Semantics models (CDSMs) bridge the gap by incorporating the principle of compositionality, which asserts that *the meaning of a complex expression is determined by the meanings of its parts and the rules governing their combination* [64]. This extends semantic modelling beyond individual words to phrases and sentences, ensuring alignment between meaning and syntactic structure.

2.2.1 Algebraic Composition

Early compositional approaches used simple mathematical operations to combine word vectors into multi-word representations. Mitchell and Lapata [13] introduced two key approaches in this area: the additive and multiplicative compositions.

Additive: The additive model combines the meanings of two components by summing their vectors, for instance:

$$\overrightarrow{fast\ food} = \overrightarrow{fast} + \overrightarrow{food}$$

This approach is computationally efficient and captures general semantic relationships in simple compositions. However, its commutative nature prevents it from distinguishing between phrases like *fast food* and *food fast*, as the order of addition does not affect the result. Additionally, additive models often *blend* the meanings of components, leading to ambiguous representations in more complex phrases.

Multiplicative: The multiplicative model, on the other hand, combines word vectors through element-wise multiplication to emphasize shared features, for example:

$$\overrightarrow{fast\ food} = \overrightarrow{fast} \odot \overrightarrow{food}$$

While this model captures intersections between components effectively, it shares the commutative limitation of the additive model. Furthermore, its filtering effect, where non-overlapping features are zeroed out—can result in overly sparse representations.

2.2.2 Tensor Based Models

To address the limitations of previous models, such as commutativity and their inability to encode syntactic and relational roles, tensor-based models were proposed. Composition methods using tensor products were first introduced by Smolensky [18] and later refined by Clark and Pulman [19]. In general, these methods address commutativity by encoding grammatical structure through non-commutative tensor products, which combine word meanings into higher-dimensional spaces. For instance, the tensor product of two vectors $\vec{w}_1 \in \mathbb{R}^m$ and $\vec{w}_2 \in \mathbb{R}^n$ is defined as:

$$\vec{w}_1 \otimes \vec{w}_2 = \begin{bmatrix} w_{11} \cdot w_{21} & w_{11} \cdot w_{22} & \dots & w_{11} \cdot w_{2n} \\ w_{12} \cdot w_{21} & w_{12} \cdot w_{22} & \dots & w_{12} \cdot w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1m} \cdot w_{21} & w_{1m} \cdot w_{22} & \dots & w_{1m} \cdot w_{2n} \end{bmatrix}$$

This operation creates a composite representation in a higher-dimensional space $\mathbb{R}^{m \cdot n}$. However, while tensor product models succeed in preserving word order and encoding grammatical relationships, they suffer from exponential growth in dimensionality as more words are combined, making them computationally inefficient for larger phrases or sentences. Second, the resulting vectors from tensor-based compositions for sentences of different lengths or structures could not be directly compared because they lived in different vector space.

Several models were proposed to mitigate these issues. Plate's Holographic Reduced Representations (1991) [65], attempted dimensionality reduction by encoding high-dimensional tensors into smaller spaces using circular convolution. However, this introduced noise and information loss, limiting the reliability of these models for precise semantic representation. Other notable efforts included structured vector space models by Erk & Padó [66] and dependency-based approaches by Clark & Pulman [19]. These models integrated syntactic information into word embeddings through dependency relations or other linguistic structures. While they added valuable syntactic sensitivity, they lacked a unified framework for composing meanings across varied sentence structures.

2.2.3 DisCoCat

Coeke et al. [20] made an interesting observation that syntactic structures, such as those derived from pregroup grammars, and semantic vector spaces share a common foundation within category theory. By modelling both as compact closed categories, they established a unified approach where syntactic information directs the compositional process of semantic meanings. In this framework, the meaning of a sentence emerges from tensor contraction operations that integrate the meanings of individual words while respecting their grammatical roles. This framework is discussed in further detail in Section 2.6.

Adjective-Noun Composition: One of the earliest works that align closely with this framework (although implemented independently) is by Baroni and Zamparelli [21]. They developed a model specifically focused on adjective-noun constructions, which follows a similar principle of type-driven composition. In their approach, adjectives are treated as linear transformations (matrices) that operate on noun vectors, functioning as mappings that modify the meaning of a noun to produce the meaning of the adjective-noun phrase, e.g., $f_{\text{bright}} : \textit{star} \mapsto \textit{bright\_star}$. Mathematically:

$$\overrightarrow{\textit{bright star}} = \overline{\textit{bright}} \times \overrightarrow{\textit{star}}$$

where $\overline{\textit{bright}}$ is the matrix representation of the adjective *bright*, $\overrightarrow{\textit{star}}$ is the vector representation of the noun *star*, and $\times$ denotes matrix-vector multiplication. The adjective matrix is trained using linear regression to approximate the semantic vector of the resulting phrase based on observed data.

In parallel, Guevara [67] proposed a regression-based approach for adjective-noun composition using Partial Least Squares Regression (PLSR) to learn data-driven transformations directly from corpus data. Unlike Baroni and Zamparelli's [21] fixed transformations, this model flexibly captures both adjective- and noun-specific contributions but depends on high-quality training data for accurate mappings.

Another influential model for adjective-noun composition is proposed by Maillard and Clark [22]. They present a tensor-based skip-gram model for learning

adjective meanings, building on the compositional framework of Coecke et al. [20] and Maillard et al. [40]. The authors extend the traditional skip-gram model [60] by representing adjectives as matrices and nouns as vectors, training these embeddings using a two-stage process. Noun vectors are first learned via negative sampling, followed by the optimization of adjective matrices. The model demonstrates competitive performance in adjective and adjective-noun similarity tasks.

Other Compositions: The empirical implementation of Coecke et al. [20] was developed by Grefenstette et al. [68], who introduced methods to construct sentence vector spaces using tensor products. This implementation was further validated in a series of papers. Grefenstette and Sadrzadeh [69] demonstrated its effectiveness for transitive and intransitive sentences. Later they improved semantic disambiguation for transitive verbs by introducing an alternative method to compute verb tensor representations, where the verb tensor is calculated as the the Kronecker product of the verb's vector representation with itself [70]. However, these composition methods come with a limtation. i.e., inability to handle vectors with negative values. To overcome these limitations, Grefenstette et al. [23] introduced a multi-step regression approach for learning transitive verb tensors. This method first learns matrices for verb phrases that approximate corpus-based sentence vectors when multiplied with subject vectors, followed by learning third-order verb tensors that, when multiplied with object vectors, reproduce verb phrase matrices. In the following years, the focus shifted toward addressing more complex linguistic phenomena. For instance, Wijnholds and Sadrzadeh [71] explored verb phrase ellipsis with tensor-based and non-linear composition, achieving improvements in tasks involving elliptical sentences. Building on this, Wijnholds et al. [24] introduced multilinear skip-gram models that leverage grammatical types for representation learning, demonstrating competitive performance against neural encoders like BERT.

2.2.4 Neural Network-Based Composition

Neural network-based approaches use deep learning architectures to construct phrase and sentence embeddings by hierarchically combining word-level representations. While some models incorporate syntactic structures explicitly, others mainly focus

on semantic relationships, deriving high-level abstractions from raw data.

Recursive Neural Networks (RecNNs): A notable example of syntactically informed composition is RecNNs, as demonstrated by Socher et al. [15]. These models use recursive structures based on parse trees to combine word vectors hierarchically, creating compositional representations for phrases and sentences. At each node of the tree, pairs of word or phrase embeddings are merged using a learned transformation function, often followed by non-linear activations like *tanh*, to produce higher-level representations. RecNNs align with syntactic structures both the compositional semantics and the hierarchical relationships, excelling in tasks like paraphrase detection and sentiment analysis.

Sentence Encoders: These models aim to generate fixed-size embeddings to capture semantic meaning of entire sentences rather than individual words, typically trained on tasks where capturing sentence-level relationships is essential. They do not explicitly encode syntactic structures but focus on learning robust semantic patterns from data. Some prominent models include:

1. *InferSent* [16]: is specifically designed for natural language inference (NLI) tasks, utilises a supervised learning approach to produce fixed-size embeddings for sentences that effectively capture semantic information. Built on Long Short-Term Memory (LSTM) networks, InferSent is trained on datasets like the Stanford Natural Language Inference (SNLI) corpus, enabling it to learn rich representations for understanding sentence relationships.

2. *Universal Sentence Encoder (USE)* [72]: pre-trained on diverse text data, generates fixed-size sentence embeddings using either a transformer architecture or a Deep Averaging Network (DAN), enabling it to capture semantic meaning and contextual relationships.

Contextualized Models: Sentence embeddings dynamically adapt to sentence-specific contexts, offering greater flexibility than static embeddings, with Sentence-BERT (SBERT) being a key advancement in sentence-level understanding.

1. *SBERT* [17]: extends the BERT architecture by utilising a Siamese network design, which allows it to process sentence pairs simultaneously and generate fixed-size embeddings. To derive a single vector representation from the token embeddings, SBERT employs pooling strategies like mean pooling, which averages the embeddings, and max pooling, which selects the maximum values across dimensions. This approach enables SBERT to produce high-quality sentence embeddings that can be fine-tuned for specific applications, significantly enhancing various Natural Language Processing (NLP) tasks such as sentence similarity, semantic search, and clustering.

2. *Other models*: like RoBERTa [73], ALBERT [74], and T5 [75] have emerged after BERT, each offering unique enhancements in pre-training strategies, efficiency, and context handling. For example, RoBERTa improves BERT by using dynamic masking and more training data, ALBERT reduces model size through parameter sharing, and T5 reframes all NLP tasks as text-to-text problems. While these models achieve strong performance on many benchmarks, they are not explicitly designed to model syntactic or grammatical structure and often rely on large-scale data rather than linguistic priors to capture such information.

   Transformer architectures have also been adapted for audio and multimodal processing. Models such as AST (Audio Spectrogram Transformer) [76] and PaSST [77] apply Transformer encoders directly to spectrogram patches, enabling long-range temporal modelling in audio classification. CLAP [78] and AudioCLIP [79] extend the CLIP framework to align text and audio embeddings, while models like AudioLM [80] integrate speech, audio, and text for generative tasks. These Transformer-based models inherently perform a form of composition through self-attention, which captures relationships between elements in a sequence and fuses multimodal features. This provides some degree of compositional capability, but it is learned implicitly and is not guided by explicit operators for combining the meanings of smaller units into larger structures. As a result, they may capture co-occurrence patterns and holistic associations without fully modelling how the meaning of a phrase emerges from its parts. Compositional

distributional models address this gap by introducing mathematically defined composition functions, enabling more interpretable and systematic handling of linguistic structure, which can be particularly important for tasks requiring fine-grained semantic reasoning over multimodal inputs.

2.3 Multimodality in Language

Purely textual models may fall short of human-like semantic understanding, as they are not grounded in perceptual modalities [26]. This has led to proposals for multimodal approaches that incorporate sensory inputs to enrich semantic representations and better connect language with real-world context. This relates to the *symbol grounding problem* [81], where symbols acquire meaning through their link to sensory and environmental experiences. For example, the word *car* is understood not only through text but also via associated visual and auditory cues, such as its shape or engine sound. Multimodal distributional semantics (MDS) aim to address this by incorporating multiple modalities—such as text, audio, and vision—to enable more grounded and context-aware representations.

2.3.1 Multimodal Distributional Semantics

Feng and Lapata [27] introduced the first multimodal distributional semantic model, leveraging a generative probabilistic framework that integrates textual and visual features from a mixed-media corpus. By representing words through distributions over latent multimodal dimensions, they demonstrated that incorporating visual information improves performance on semantic similarity tasks compared to text-only models. Although their results showed gains in correlation with human judgments, the performance remained below the state-of-the-art benchmarks at that time. Building on this foundation, Silberer and Lapata [28] introduced grounded models of semantic representation, where visual information was used to enrich textual meaning. They demonstrated that grounding textual concepts in visual features could improve semantic similarity and relatedness, particularly for words whose meanings are closely tied to real-world objects, such as *cat* or

car. Similarly, Bruni et al. [29] proposed a framework where visual features extracted from images were combined with textual embeddings. They explored early fusion and late fusion techniques to improve semantic similarity across modalities. Further work, such as Lazaridou et al. [30], proposed a grounded multimodal Skip-gram model, which jointly learned word embeddings and visual features, allowing for a more unified approach.

Building on the success of vision-based approaches, researchers began integrating auditory modalities into multimodal distributional models. Early work by Kiela & Clark [31] and Lopopolo & Miltenburg [32] demonstrated that sound, too, could significantly ground word meanings. The work of Kiela & Clark [33] further extended the idea by incorporating audio alongside text, enabling models to evaluate semantic similarity based on both auditory and textual features. They demonstrated that audio features, when combined with text, could provide richer semantic grounding for sound-related concepts like *rain* and *guitar*, where auditory cues are key to understanding. Inspired by vision-based models, they introduced a neural audio embeddings (NAE) model, which outperformed traditional methods like bag-of-audio-words (BoAW). The learnt embeddings were evaluated on an auditory variant of MEN dataset [11] called AMEN [33]. This work inspired extensions to other neural architectures, document-level tasks, and multimodal applications, demonstrating the potential of auditory grounding to enrich semantic understanding [34–36].

2.3.2 Multimodal Compositional Distributional Semantics

The exploration of compositional multimodal distributional semantics has recently gained momentum, with the focus shifting from words to more complex linguistic structures like phrases and sentences. Building on the success of multimodal image-based models for distributional semantics, researchers are increasingly investigating composition models that integrate textual and visual modalities. Recently, Lewis et al. [37] evaluated the compositional capabilities of vision-language models, particularly CLIP [39], in combining linguistic and

visual modalities. Their work benchmarked CLIP and compositional distributional semantic models on tasks requiring binding of concepts in single-object, two-object, and relational contexts. CLIP performed well in single-object tasks, but its ability to handle abstract compositionality and variable binding proved inadequate, revealing critical limitations and emphasizing the need for more sophisticated multimodal representations. Building on this, the work of Wazni et al. [38] introduces VerbCLIP, a model specifically designed to improve the representation of verbs in vision-language frameworks. By integrating CDSMs into CLIP's structure, this framework leverages tensor-based methods to capture the roles of verbs alongside their associated subjects and objects. VerbCLIP demonstrates its strengths in capturing syntactic and semantic structures, outperforming CLIP in tasks such as verb disambiguation and scene understanding across multiple datasets.

The literature on multimodal learning in language reveals several critical insights. The vision-based models have made significant progress in integrating visual and textual modalities, particularly in capturing compositional semantics for phrases and relational structures. However, despite these advancements, the integration of auditory information within a compositional framework remains an under-explored area. This thesis seeks to address this gap by incorporating auditory information to enhance semantic grounding for concepts inherently tied to sound, such as *thunder*, *rain*, and *guitar*, which cannot be fully represented through text or visual data alone.

2.3.3 Fusion Techniques

Integrating linguistic and perceptual cues involves multimodal fusion techniques that combine information from different modalities into a cohesive representation. These techniques can be categorised as early, middle, and late fusion [29, 33].

(a) *Early fusion* sometimes referred as joint learning, integrates multiple modalities by optimizing a shared objective, ensuring aligned and semantically enriched representations. This is inspired by human cognition, where sen-

sory inputs are processed together to form coherent perceptions.

(b) *Middle fusion* involves independently learning representations for each modality and combining them into a unified representation before computing the final scores. The representations are typically combined using methods like concatenation or weighted addition.

(c) *Late fusion* processes each modality independently through separate models and combines their outputs at the decision-making stage. For example, to compute similarity scores, each modality would generate its score independently, and these scores would then be aggregated.

Lazaridou et al. [30] utilised *early fusion* in their multimodal skip-gram model, seamlessly integrating visual features with linguistic contexts. By jointly predicting word contexts and aligning visual representations with word embeddings, their approach effectively captured visual-linguistic relationships, significantly enhancing semantic understanding. In contrast, *middle* and *late fusion* methods provide flexibility by enabling each modality to be optimised independently with tailored training objectives. For instance, Bruni et al. [29] employed *early fusion* by concatenating text and visual embeddings to create multimodal representations and implemented *late fusion* by combining similarity scores derived independently from textual and visual embeddings to assess semantic relatedness. They demonstrated the value of these techniques using concept pairs such as *cat* and *dog* or *car* and *road*, highlighting how visual grounding complements text-based distributional semantics. On the other hand, Kiela and Clark [33] incorporated auditory information into multimodal fusion. They applied *middle fusion* by combining textual and auditory embeddings through weighted concatenation and *late fusion* by aggregating similarity scores using weighted averaging.

2.4 Audio Learning

Traditionally, audio processing aims to extract meaningful features from raw signals using domain-specific knowledge. Classical methods relied on human

perception models to identify key features, such as:

(a) *Mel-Frequency Cepstral Coefficients (MFCCs) [82]*: capture spectral features while approximating the human ear's sensitivity to different frequencies. Widely used in speech recognition, these features excel in controlled environments but struggle with noisy or complex audio data.

(b) *Chroma Features [83]*: represent the harmonic content of music, became the standard for music information retrieval (MIR) tasks. These features capture the tonal qualities of music and were particularly useful in classifying musical genres, analyzing mood, and identifying key elements of music.

While effective for their time, these handcrafted features lacked the flexibility to handle diverse datasets and failed to capture high-level abstract relationships, such as the complex interplay of harmonics in polyphonic music or the subtle distinctions between overlapping environmental sounds.

The limitation led to the adoption of deep learning techniques, particularly using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which enable automatic learning of hierarchical features directly from raw audio. One of the earliest effort applied CNNs to music classification and tagging, learning features directly from raw audio waveforms and spectrograms [84]. Following the trend, Piczak [85] used CNNs to classify environmental sounds like sirens and animal noises by learning features from log-mel spectrograms, and Graves et al. [86] applied LSTM networks, to speech recognition, capturing long-range temporal dependencies and improving accuracy.

Inspired by the progress in image models (e.g., AlexNet [87], ResNet [88]), researchers sought to develop audio representations that could generalise across tasks, whether environmental, musical, or speech. Self-supervised and unsupervised learning approaches emerged, allowing models to learn audio features without the need for extensive labeled datasets. Aytar et al. [89] laid the foundation and presented *SoundNet*, a weakly supervised deep convolutional network that leveraged the audiovisual synchronization in videos to learn rich audio

representations. Other popular models include, but are not limited to:

(a) *VGGish [90]*: The VGGish embeddings were first introduced by Hershey et al. [90], where a modified VGG CNN [91] was trained using mel-spectrograms as input. Later, Google released a variant[2] pre-trained in a supervised manner on the weakly labeled YouTube-8M dataset [92], which contained overlapping tags. This dataset was later replaced with the more robust AudioSet [93] with 2 million audio clips labeled with 527 tags, for improved label accuracy and better representation of audible events. These 128-dimensional embeddings provide powerful features for tasks such as audio classification and retrieval.

(b) *Kumar [94]*: embeddings are generated using a supervised CNN based on a VGG-like architecture [91] with mel spectrograms as input. These embeddings are pre-trained in a supervised way on a subset of AudioSet [93] that includes approximately 22,000 clips from YouTube videos across 527 sound categories. The resulting embeddings have a 1024-dimensional feature representation, appropriate for general purpose audio tasks.

(c) *OpenL3:* OpenL3 [95] is a self-supervised method for generating audio feature representations, extending the L3-Net architecture [96]. The model is available in multiple configurations, with one pre-trained on a music subset and the other on an environmental sound subset of AudioSet [93], containing 296K and 195K YouTube videos, respectively. During training,

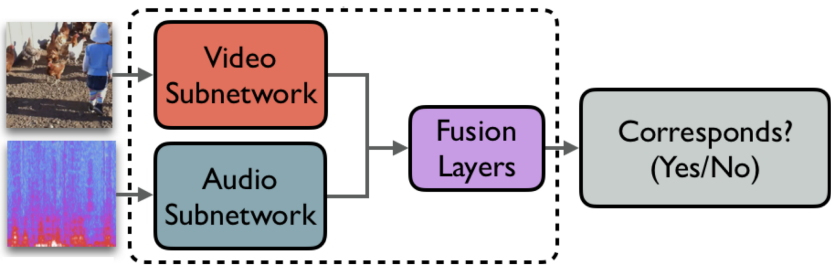


Figure 2.1: OpenL3 Architecture.

[2]VGGish: `https://github.com/tensorflow/models/tree/master/research/audioset/vggish`

the model learns to determine whether the audio corresponds to the visual content in the video or not. After training, only the audio subnetwork is required to extract embeddings directly from the audio data (Figure 2.1). The resulting embeddings are represented as 512-dimensional feature vectors. This self-supervised learning approach enables the embeddings to generalise well across a variety of audio-related tasks, including sound event detection and audio classification[3].

(d) *Other models*: popular pretrained audio models, such as Speech2Vec [98], Wav2Vec [99], and Audio ALBERT [100], are primarily designed for speech-related tasks like automatic speech recognition (ASR) and spoken language understanding. While these models are effective for processing spoken language, they are less suited for non-speech audio, such as environmental or musical.

2.5 Evaluation Benchmarks

This section reviews widely used datasets for textual and multimodal evaluation benchmarks, highlighting key gaps in the existing literature.

2.5.1 Textual Benchmarks

Table 2.1 summarises common textual benchmarks for evaluating semantic similarity across various linguistic levels, including words, phrases, and sentences.

Table 2.1: Popular datasets for textual evaluation across word, phrase, and sentence levels.

Dataset	Category	Level	Task	Size	Examples
WordSim353 [10]	Nouns	Word	Semantic Similarity	353 pairs	*day–summer, movie–star*
RG65 [4]	Nouns, Adjectives	Word	Semantic Similarity	65 pairs	*food–fruit, coast–hill*
MEN [29]	Nouns, Adjectives, Verbs	Word	Semantic Similarity	3,000 pairs	*bed–sleep, fell–love*
SimLex999 [9]	Nouns, Adjectives, Verbs	Word	Semantic Similarity	999 pairs	*smart–dumb, fast–rapid*
Mitchell & Lapata [41]	Adjective-Nouns	Phrase	Semantic Similarity	108 pairs	*dark eye–left arm*
Vecchi [42]	Adjective-Nouns	Phrase	Plausibility	28K phrases	*rear liver–funny juice*
Mitchell & Lapata [13]	Subject-Verb	Sentence	Semantic Similarity	200 pairs	*The fire glowed–burned*

For **word-level** tasks, datasets like WordSim353 [10] and RG65 [4] focus on semantic similarity for nouns and adjectives, with examples such as *day – summer*

[3]Liu et. al. [97] present a comprehensive survey on the auditory self supervised learning methods.

and *food – fruit*. Larger datasets like MEN [29] and SimLex999 [9] extend these evaluations to include verbs, enabling models to capture a broader range of semantic relationships. At the **phrase-level**, benchmarks such as Mitchell & Lapata [41] test semantic similarity, as seen in examples like *dark eye – left arm*. Additionally, Vecchi [42] introduces a larger dataset for plausibility, distinguishing between realistic combinations (*rear liver*) and implausible ones (*funny juice*). Finally, for **sentence-level** evaluations, Mitchell & Lapata [13] provide 200 subject-verb pairs focused on semantic similarity, such as *The fire glowed – burned*. These datasets offer a range of granularities, allowing for comprehensive evaluation of compositional and contextual semantics across linguistic levels.

2.5.2 Audio/Multimodal Benchmarks

Table 2.2 provides an overview of popular multimodal benchmarks used for word- and sentence-level evaluations.

Table 2.2: Multimodal benchmarks for word and sentence level evaluations.

Dataset	Modality	Level	Task	Size	Examples
UrbanSound8K [101]	Audio	Word	Audio Classification	8,732 clips	The word *Footsteps* with multiple sounds.
ESC-50 [102]	Audio	Word	Audio Classification	2K clips	The word *Dog* with 50 sounds.
SemSim/VisSim [28]	Text, Images	Word	Semantic & Visual Similarity	7576 pairs	ant – rat, axe – pin
AudioCaps [43]	Text, Audio	Sentence	Audio Captioning	51K captions	ID, Caption (1, *Rain is falling continuously*)
Clotho [44]	Text, Audio	Sentence	Audio Captioning	5K audio files	Audio samples with five captions each.
AudioSet [93]	Text, Image, Audio	Sentence	Audio Classification	2M clips	*barking* is annotated as *Animal*, *Pets*, and *Dog*.

These datasets span multiple modalities, including audio, text, and images, and are designed for tasks such as audio classification, captioning, and semantic similarity. At the **word-level**, benchmarks like UrbanSound8K [101] and ESC-50 [102] focus on audio classification, testing models' ability to distinguish sounds associated with specific words. Examples include *Footsteps* and *Dog*, each paired with multiple audio variations. Similarly, SemSim/VisSim [28] evaluates semantic and visual similarity, leveraging text and image pairs such as *ant – rat* and *axe – pin*. For **sentence-level** tasks, datasets like AudioCaps [43] and Clotho [44], assess audio captioning models by linking audio inputs to textual descriptions. For instance, captions like *Rain is falling continuously* provide

context for audio samples. AudioSet [93] extends to text, image, and audio modalities, focusing on audio classification with rich annotations such as *barking* labeled as *Animal, Pets*, and *Dog*.

Scarcity of Audio Phrase Datasets:

Existing benchmarks are limited in scope. *Word similarity* datasets such as MEN [29] and WordSim353 [10] lack sound-relevant adjective pairs; for example, WordSim353 includes *smart–stupid* and *Japanese–American*, which are not audio- or sensory-related. This gap extends to *phrase similarity* benchmarks like Mitchell & Lapata [41], which focus on frequent English adjectives applied to many nouns, with minimal auditory overlap. Existing *multimodal datasets* such as AudioCaps [43] and Clotho [44] link full sentences to audio captions but lack fine-grained evaluations for phrases such as adjective–noun combinations. This thesis addresses these gaps by introducing a new multimodal adjective–noun phrase similarity dataset in Chapter 4.

2.6 Categorical Models and Tensor-Based Semantics

The Categorical Compositional Distributional Semantics framework (DisCoCat) introduced by Coecke et al. [20] unifies syntactic structures and semantic representations using category theory, mapping pregroup grammars and vector spaces to a shared compact closed category. This enables a systematic composition of meanings for linguistic units by combining tensors and linear maps with grammatical reductions. This section briefly discusses the DisCoCat model and its extension to Combinatory Categorial Grammar (CCG) by Maillard et al. [40].

2.6.1 Categories

Category theory is a branch of mathematics that provides a highly abstract framework for studying structures (called *objects*, such as words or phrases) and the relationships between them (called *morphisms*, e.g., functions or mappings).

It is governed by a set of fundamental principles that define the behavior of objects and morphisms. These principles include:

- *Composition of Morphisms (functions)*: follows the associative rule: if $f : A \to B$ and $g : B \to C$, then their composition $g \circ f : A \to C$ satisfies: $(f \circ g) \circ h = f \circ (g \circ h)$.
- *Identity Morphisms*: Each object A has an associated identity morphism, denoted as $\mathrm{I}_A : A \to A$, which acts as a neutral element under composition. For any morphism $f : A \to B$, the identities hold: $\mathrm{I}_A \circ f = f \quad \text{and} \quad g \circ \mathrm{I}_A = g$.

Monoidal Categories: extend these principles by introducing additional structure. Specifically, monoidal categories define a tensor product ($\otimes$) and a unit object (I) that generalise how objects and morphisms interact. This extension allows for richer modelling of systems where multiple objects combine in parallel. Key additional properties include:

- For any objects A, B, C, $(A \otimes B) \otimes C \cong A \otimes (B \otimes C)$.
- For any object A, $A \otimes I \cong A \cong I \otimes A$.

For instance, in the category of sets, the tensor product can be represented by the Cartesian product of sets, where the unit object is a singleton set. For two sets $A = \{a, b\}$ and $B = \{1, 2\}$, the tensor product $A \otimes B$ is the set of ordered pairs:

$$A \otimes B = \{(a,1), (a,2), (b,1), (b,2)\}.$$

In *symmetric monoidal category*, the tensor product is commutative, formalised by isomorphism, meaning: $A \otimes B \cong B \otimes A$.

Compact Closed: A monoidal category is said to be compact closed if every object A has a *left adjoint* A^l and a *right adjoint* A^r. These adjoints are connected to A through special morphisms:

$$\eta^l : I \to A \otimes A^l, \quad \eta^r : I \to A^r \otimes A$$

$$\varepsilon^l : A^l \otimes A \to I, \quad \varepsilon^r : A \otimes A^r \to I$$

For a *symmetric compact closed category*, the left and right adjoints of each object collapse into one, so that: $A^* = A^l = A^r$.

One significant example of a compact closed category is **FVect**, the category of finite-dimensional vector spaces and linear maps. In **FVect**, vector spaces represent semantic meanings of words, while linear maps capture compositional rules that combine these meanings. This structure aligns seamlessly with pregroup grammar, where syntactic reductions are represented as morphisms.

The interaction between **FVect** and pregroup grammar is formalised in the product category $\mathbf{FVect} \times P$, where P corresponds to syntactic types and reductions. This allows syntactic structure to guide the semantic composition of vectors. For instance, an adjective type $n^l \cdot n$ is represented in **FVect** as a matrix that acts on noun vectors to create a noun phrase.

2.6.2 Pregroup Grammar

Introduced by Lambek [103], pregroup grammar is a type-logical grammar that provides a mathematical framework for modelling the syntactic structure of natural language. It is based on the *partially ordered monoid* with a unit element (denoted as 1), where each element has left and right adjoints (inverses). These adjoints allow for the reduction of grammatical types, enabling syntactic parsing. Formally, for any element A in the pregroup, the following properties hold:

$$A \cdot A^l \leq 1 \quad \text{and} \quad A^r \cdot A \leq 1$$

where A^l and A^r are the left and right adjoints of A, and 1 is the identity element. These adjoints allow for type cancellation, simplifying complex type sequences into valid grammatical structures.

Each word in a sentence is assigned a grammatical type, derived from *atomic* types (e.g., n for nouns, s for sentences). *Complex types* are constructed from

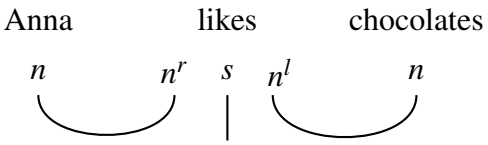
the atomic types using the monoidal product ($\cdot$) and the adjoints (A^l, A^r), which encode the syntactic dependencies between words. For instance:

- A *noun* is represented by the atomic type n.
- An *adjective* modifies a noun and is represented by the type $n^l \cdot n$, indicating it consumes a noun to the right and results in a noun phrase.
- A *transitive verb* expects a noun phrase on its left (subject) and another on its right (object) to form a sentence. Its type is $n^r \cdot s \cdot n^l$.

Words in a sentence combine according to reduction rules based on their grammatical types. These rules are applied sequentially to simplify the sentence into its grammatical form. Reduction diagrams, often referred to as *wire diagrams*, visually represent the type reduction process.

Example 1: A transitive sentence *Anna likes chocolates* is analyzed using pregroup grammar by assigning types to each word: n for the noun *Anna* (subject) and *chocolate* (object), and $n^r \cdot s \cdot n^l$ for the transitive verb *likes*.

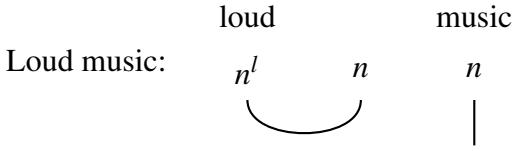
Anna likes chocolates:



Type Sequence: $n \cdot (n^r \cdot s \cdot n^l) \cdot n \to n \cdot n^r \cdot s \cdot n^l \cdot n \to s \cdot n^l \cdot n \to s.$

The reductions $n \cdot n^r \to 1$ and $n^l \cdot n \to 1$ simplify the structure to the sentence type s. First, the subject (n) and the right adjoint of the verb (n^r) combine, reducing $n \cdot n^r \to 1$. This simplifies to $s \cdot n^l \cdot n$, where s represents the sentence's core type. Next, the left adjoint of the verb (n^l) and the object noun (n) combine, reducing $n^l \cdot n \to 1$. The remaining type s confirms the grammatical validity of the sentence.

Example 2: For an adjective-noun phrase say *loud music*, the reduction is as follows:



Type Sequence: $(n^l \cdot n) \cdot n \to 1 \cdot n \to n.$

The reduction $n^l \cdot n \to 1$ simplifies to the type n, representing a noun phrase. In this example, the adjective "loud" is assigned the type $n^l \cdot n$, indicating it modifies a noun. The noun "music" is assigned the atomic type n. During the reduction, the left adjoint n^l of the adjective combines with the noun n to produce 1, leaving n as the resulting type. This final type n confirms the grammaticality of the phrase "loud music" as a well-formed noun phrase.

Coecke et al. [20] leverages the compact closed structure of both **FVect** and pregroup grammar to systematically compute the meanings of phrases and sentences. Syntactic structures derived from pregroup grammars guide tensor-based operations in **FVect**, ensuring that grammatical dependencies are preserved in semantic compositions. For example, in the DisCoCat framework, a transitive verb is represented as a third-order tensor that operates on subject and object noun vectors to produce a sentence vector:

$$\overrightarrow{sentence} = \mathbf{T} \cdot (\overrightarrow{subject} \otimes \overrightarrow{object})$$

where $\mathbf{T}$ is the tensor representation of the verb, and $\overrightarrow{subject}$, $\overrightarrow{object}$ are the subject and object vectors. This mapping of grammatical structure to tensor-based operations ensures that the meaning of complex linguistic units can be derived compositionally from their constituents.

2.6.3 CCG and Tensor-based Semantics

Building on the DisCoCat framework, Maillard et al. (2014) proposed an extension of tensor-based semantics to Combinatory Categorial Grammar (CCG). CCG provides a more flexible syntactic formalism compared to pregroup grammars, al-

lowing for richer linguistic constructs such as type-raising and composition. This extension highlights the compatibility of tensor-based distributional semantics with CCG's combinatory rules. In this section, first we will discuss briefly about CCG and then about its extension to tensor based semantics.

Combinatory Categorial Grammar (CCG): developed by Mark Steedman [104] is designed to represent both syntax and semantics in a unified framework. Similar to the pregroup grammar, CCG categorises linguistic elements into basic and complex categories. *Basic categories*, such as S for sentences and NP for noun phrases, represent fundamental grammatical types. *Complex categories*, constructed using these basic categories and slashes ($/$ and $\backslash$), define functions that describe how words combine with others. These slashes indicate the direction of combination: $/$ represents a function expecting an argument to its right, while $\backslash$ expects an argument to its left. Complex categories, therefore, serve as functional operators, taking arguments and producing resulting categories.

To combine categories, CCG relies on a set of combinatory rules that define how syntactic categories interact. These rules enable the composition of more complex grammatical structures from basic categories. This thesis discusses only two rules, i.e, Forward application and Backward application.

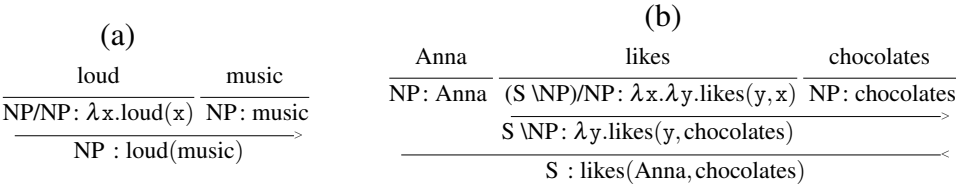


Figure 2.2: CCG derivations demonstrating (a) adjective-noun composition and (b) subject-verb-object sentence composition, using forward and backward application respectively, and incorporating semantic representations.

Forward application ($>$) applies when a function category with the form X/Y (a function expecting an argument of type Y) is followed by a category Y on its right, to produce the result category X. For example, consider the phrase *loud music*. The word *loud* is represented as NP/NP, meaning it modifies a noun phrase, and *music* is NP, a noun phrase. Using Forward Application, NP/NP

combines with *NP* to produce *NP*, resulting in the complete phrase *loud music* (Figure 2.2 (a)).

Backward application (<) applies when a function category $X\backslash Y$ (expecting an argument of type Y) is preceded by a category Y on its left. The result of this combination is X, the outcome of the function applied to its argument. For example, in the sentence *Anna likes chocolates*, the word *likes* is represented as $(S\backslash NP)/NP$, a transitive verb that expects two arguments: an object (*NP*) on its right and a subject (*NP*) on its left. First, it combines with *chocolates* (*NP*) using Forward Application, resulting in $S\backslash NP$. Then, this intermediate result combines with *Anna* (*NP*) using Backward Application, yielding *S*, a complete sentence (Figure 2.2 (b)).

Semantic Integration: CCG integrates both syntax and semantics by associating each syntactic category with a corresponding semantic type, allowing combinatory rules to operate uniformly on both syntactic and semantic levels. Lambda calculus is a widely used formalism for encoding these semantic representations [104].

Figure 2.2 (a) demonstrates the adjective-noun composition where the adjective *loud* is represented as a function $\lambda x.\text{loud}(x)$, modifying the noun *music* to produce the semantic interpretation $\text{loud}(\text{music})$. Figure 2.2 (b) illustrates a subject-verb-object sentence composition, where the transitive verb *likes* is represented as a higher-order function $\lambda x.\lambda y.\text{likes}(y,x)$. The subject *Anna* and the object *chocolates* are sequentially applied to this function through forward and backward application rules, yielding the complete semantic representation $\text{likes}(\text{Anna},\text{chocolates})$.

Tensor-Based Semantics: Tensor-based semantics for CCG, introduced by Maillard et al. [40], interprets grammatical types as tensor spaces and implements CCG's combinatory rules through tensor contraction. In this framework, words are represented as tensors of varying orders based on their grammatical and functional roles. Unary functions, such as adjectives and intransitive verbs, are

modeled as second-order tensors (matrices) that operate on first-order tensors (vectors) like nouns to produce refined representations. For example, an adjective like *fast* modifies a noun like *food* through matrix-vector multiplication, capturing the adjective's role as an operator that refines the noun's semantic properties. Binary functions, such as transitive verbs, require two arguments (subject and object) and are modeled as third-order tensors, while ternary functions, such as ditransitive verbs like *gives*, are represented as fourth-order tensors, enabling the hierarchical representation of complex sentence structures. This framework integrates syntactic and semantic compositionality by treating adjectives, verbs, and other functional words as linear maps or tensors that transform vector representations of their arguments. For instance, an adjective is represented as a matrix M_{adj} that operates on the vector of a noun ($\overrightarrow{noun}$) to produce the vector representation of the resulting adjective-noun phrase:

$$\overrightarrow{phrase} = M_{\text{adj}} \times \overrightarrow{noun}$$

Also, transitive verbs are modeled as third-order tensors T_{verb}, capable of combining subject and object vectors into a sentence-level vector. The representation is computed as:

$$\overrightarrow{sentence} = T_{\text{verb}} \cdot (\overrightarrow{subj} \otimes \overrightarrow{obj})$$

Here, $\otimes$ denotes the Kronecker product, which combines the subject ($\overrightarrow{subj}$) and object ($\overrightarrow{obj}$) vectors into a higher-dimensional space.

2.7 Conclusion

The review of the literature on compositional distributional semantics reveals that these methods have significantly advanced language representation over the years. However, this is still a debate as to what extent these models ground the perceptions as humans naturally do. This is because, one of the key limitations of purely textual models is their inability to account for the fact that human language is inherently grounded in sensory and perceptual experiences. If the ultimate

goal of language systems is to achieve human-like comprehension, then why not leverage the same capability for achieving human-like understanding?

The success of compositional distributional semantics, particularly when combined with the categorical framework [20], has provided deep insights into how the linguistic meaning can be systematically deduced. Prominent contributions (e.g., Baroni et al. [21], Grefenstette et al. [23], Maillard & Clark [22], Wijnholds et al. [24]) have demonstrated the potential of these models to handle phrase and sentence-level compositions effectively. However, despite these advancements, the extent to which these compositional models are truly grounded remains an open question.

Building on the success of multimodal distributional semantics (Feng and Lapata [27], Silberer and Lapata [28], Bruni et al. [29], Kiela & Clark [33]), only recently have researchers (e.g., Lewis et al. [37], Wazni et al. [38]) demonstrated that multimodal compositional distributional semantics has the potential to outperform state-of-the-art models. Moreover, the overwhelming focus of multimodal distributional semantics—whether standalone or in compositional settings—has been on vision, largely due to the abundance of resources in this domain. In contrast, the integration of auditory perceptions into these models has received limited attention, highlighting a significant gap in the literature.

This thesis aims to bridge this gap by extending the tensor-based CCG framework of Maillard et al. [40] to integrate auditory data into compositional semantics. By modelling audio signals and linguistic elements within a unified framework, it aims to evaluate the compositional meaning of phrases (e.g., adjective-nouns) in a multimodal context. Drawing inspiration from the multimodal distributional work of Kiela and Clark [33] and the compositional skip-gram model of Maillard & Clark [22], this research extends these methodologies to handle auditory phrase compositions. In the next chapter, a detailed discussion of the proposed framework is provided.

Chapter 3

Statistical Methods for *MultiCoDi*

This chapter introduces a type-driven multimodal compositional distributional framework, ***MultiCoDi****, inspired by DisCoCat [20]. The framework aims to integrate multimodal information, specifically combining auditory and textual data, to capture the distinctive properties of words and their grammatical roles for phrase-level compositions.*

In distributional semantics, words are often represented as vectors in high-dimensional spaces, capturing their meanings based on patterns in large corpora. Although such methods effectively model individual words, they often fall short when addressing *compositionality*, as discussed in Chapter 2. Take adjectives, for instance. Mitchell and Lapata [13] proposed vector addition as a simple approach to compose adjectives and nouns in distributional semantics. Later, in a series of papers [21, 22, 69], it was argued that vector addition is not appropriate for composition as it is commutative. Such behavior is problematic for capturing the hierarchical and directional nature of language. Additionally, adjectives serve a modifying role, transforming the meaning of nouns. This transformation cannot be accurately modeled by simple addition, which treats both components symmetrically and ignores their syntactic roles. Instead, it necessitates the use of functional representations, such as maps. In finite dimensions, maps are approximated by matrices and adjective-noun phrase composition becomes matrix-vector multiplication, a non-commutative operation. Different methods were put forward for learning the adjective matrices; Baroni et al. [21] used linear

regression, while Maillard et al. [22] and Wijnholds & Sadrzadeh [71] developed a tensorial extension of the Skipgram model [60].

Although these methods have been widely applied to text, their extension to multimodal settings remains limited. Existing work in multimodal distributional semantics has largely focused on integrating word-level representations of text and images [29, 30] and audio [31]. While recent studies have explored matrix-based phrase composition integrating text and images [37, 38], it has never been extended to audios.

This chapter aims to fill this gap by introducing a framework called **MultiCoDi** that extends compositional distributional semantics to integrate textual and auditory data. The methodology follows compositional distributional semantics of Baroni et al. [21], parsing linguistic phrases into Combinatory Categorial Grammar (CCG) trees [105], which are then used to learn multimodal embeddings. Adjectives, in both textual and multimodal contexts, are represented as matrices, while nouns are represented as vectors. These representations are learned using various machine learning algorithms. Matrix-vector multiplication is employed to derive embeddings for adjective-noun phrases.

The chapter begins by exploring methods for obtaining auditory and textual vector embeddings. Section 3.2 explains the proposed framework to learn matrices in single and multimodal settings. Section 3.3 details the implementation of the skip-gram model to integrate auditory and textual embeddings. The chapter concludes with a discussion on how the integration of audio features with linguistic data may enhance the representation of compositional phrase meanings.

3.1 Vector Representations: Text and Audio

This section provides a brief overview of the textual and auditory embeddings, alongwith the specific pretrained embeddings utilised in this thesis. For a more comprehensive discussion, please refer to Chapter 2.

Textual Embeddings: Textual embeddings provide compact vector represen-

tations of words, essential for capturing semantic relationships. Early models like Latent Semantic Analysis (LSA) [106] used statistical patterns and techniques like Singular Value Decomposition (SVD) to create dense word vectors. Enhancements such as Pointwise Mutual Information (PMI) and Positive PMI (PPMI) [59] improved robustness, but these methods relied heavily on global co-occurrence statistics. Word2Vec [107] addressed this limitation by introducing CBOW and Skip-gram, providing context-sensitive embeddings. Subsequent advancements like GloVe [7] combined co-occurrence statistics with local context, while FastText [61] incorporated subword information to handle rare and out-of-vocabulary words. However, these models produced static embeddings, where a word's representation remained fixed regardless of context. Addressing this limitation, BERT (Bidirectional Encoder Representations from Transformers) [8] introduced contextualised embeddings, allowing a word's representation to adapt based on its surrounding context. Unlike earlier models, BERT employs a bidirectional transformer architecture, capturing both left and right contexts simultaneously. Pre-trained on large corpora, BERT bridged the gap by providing dynamic, context-aware representations, setting new benchmarks across a wide range of NLP tasks. Building on BERT, Sentence-BERT (SBERT) [17] extended its capabilities to sentence-level tasks. By fine-tuning BERT with a siamese network structure, SBERT generates fixed-size, semantically meaningful embeddings for entire sentences or phrases. These embeddings excel in tasks like semantic similarity and search due to their ability to represent sentence-level meaning effectively.

This work utilises state-of-the-art transformer-based architectures, BERT and SBERT, recognised for their exceptional performance across diverse NLP tasks. Specifically, pre-trained **BERT-base-Uncased** generates **768-dimensional** embeddings for individual words. For phrases, **Sentence-BERT (SBERT)**, an extension of BERT fine-tuned for sentence-level tasks, is employed to produce **768-dimensional** embeddings. These embeddings are derived from the hidden states of the final layer, effectively capturing rich semantic information learned

during pre-training.

Auditory Embeddings: Pre-trained audio embeddings offer compact numerical representations of audio data by capturing essential characteristics such as tone, rhythm, speech content, and acoustic features. These embeddings are learned from large-scale audio datasets and allow for efficient processing of complex audio signals in tasks like classification and retrieval. Historically, audio data was represented using handcrafted features such as MFCCs. While useful, these features often lacked the semantic depth and generalization needed across domains. *Pre-trained* models like VGGish [90], SoundNet [89], YAMNet[1], and OpenL3 [95] leverage deep learning architectures to extract more meaningful representations. Among these, OpenL3, stands out for its versatility and robustness. OpenL3 is a self-supervised model that generates 512- or 6144-dimensional audio embeddings. Using a convolutional architecture, it processes Mel-spectrograms with 256 frequency bands to extract features from raw audio.The model processes the Mel-spectrogram using a stack of convolutional layers that progressively extract relevant features from the raw audio input. It has two variants: OpenL3 (Environmental) and OpenL3 (Musical), both trained on data from the AudioSet dataset [93].

In this work, pretrained **OpenL3 (Environmental)** model was utilised to extract audio features, aligning with the dataset's characteristics, as over 80% of the data comprises environmental sounds. Ambiguous items such as *punch*, *clap*, and *whistle*, which could belong to either environmental or musical categories, were addressed through this choice. Using OpenL3's default setting of **10** frames per second, embeddings were extracted for each audio file at regular intervals, corresponding to a hop size of **0.1** seconds. A single, robust representation for each audio sample was created by averaging all frame-level embeddings, resulting in a consolidated **512-dimensional** embedding per audio.

[1] https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

3.2 MultiCoDi

The methodology is based on the grammatical types of Combinatorial Categorial Grammar (CCG, [105]), which has wide-coverage parsers [108, 109]. Syntactic types of CCG are either atomic, e.g. noun phrases: *NP*, or functional and of the form Y/X or $Y\backslash X$; depending on where they take their argument. An example is an adjective: NP/NP, which takes its argument on the right, producing an adjective-noun phrase (AN). Types are composed with each other using a set of rules, e.g. forward and backward application and composition. An example of forward application is when an adjective composes with a noun phrase, producing a noun phrase:

$$\frac{NP/NP \quad NP}{NP}>$$

We work along side the distributional semantics of CCG [40], where a word W with a functional type of n arguments is assigned $\mathbb{W}_{(n)}$, an $(n+1)$th-order tensor $\mathbb{W}_{i_1 \ldots i_{n+1}}$ in the space $V_1 \otimes \ldots \otimes V_n \otimes V_{n+1}$ with V_i's are vector spaces. Given a functional word W of n arguments and semantics $\mathbf{d}_1, \ldots, \mathbf{d}_n$ of its arguments, we denote by $\mathbb{W}_{(n)}\mathbf{d}_1 \ldots \mathbf{d}_n$ the application of the representation of W to its arguments' representations.

When W is a noun, $\mathbb{W}_{i_1}$ is a 1st-order tensor or a vector. For W an adjective, $\mathbb{W}_{i_1 i_2}$ is a 2-nd order tensor or a matrix. The objective function of a distributional algorithm that learns a vector is that of the original skipgram, the one of a tensor is described below, referred to as TSG for Tensor SkipGram:

$$\sum_{c \in C} \log \sigma(\mathbb{W}_{(1)}\mathbf{d}_1 \cdot \mathbf{c}) + \sum_{\overline{c} \in \overline{C}} \log \sigma(-\mathbb{W}_{(1)}\mathbf{d}_1 \cdot \overline{\mathbf{c}})$$

In the case of AN phrases TSG simplifies to the following, where $\mathbf{A}$ is the adjective matrix, $\boldsymbol{n}$ the vector of the noun:

$$\textbf{TSG}: \qquad \sum_{\boldsymbol{c}' \in \mathcal{C}} \log \sigma\left(\mathbf{A}\boldsymbol{n} \cdot \boldsymbol{c}'\right) + \sum_{\overline{\boldsymbol{c}}' \in \overline{\mathcal{C}}} \log \sigma\left(-\mathbf{A}\boldsymbol{n} \cdot \overline{\boldsymbol{c}}'\right) \tag{3.1}$$

3.2.1 Combining the Audio with the Text

We used two different methods for combining audio with text. In the *first* method, we concatenated their vectors (**AT-Concat**) and used the result as an input to training. In the *second* method, we trained a joint audio-text matrix (**AT-Joint**), where one representation was used as a signal to improve the other.

Linear Regression For linear regression, we trained adjective matrices $\boldsymbol{A}$ to effectively model the relationships between observed adjective-noun vectors $\boldsymbol{p}$ and noun vectors $\boldsymbol{n}$. The underlying mathematical relationship is expressed as:

$$\boldsymbol{p} = \boldsymbol{A}\boldsymbol{n} \tag{3.2}$$

In this equation, $\boldsymbol{p}$ represents the target adjective-noun vector, $\boldsymbol{A}$ is the matrix that captures the effect of the adjective on the noun, and $\boldsymbol{n}$ is the corresponding noun vector. To optimise the training process, we employed a vanilla regression technique utilizing a partial least squares (PLS) approximation.

AT-Concat Regression The AT-Concat Regression method extends the single-modality regression approach (Equation 3.2) by incorporating both audio and textual representations of nouns. This adaptation is expressed mathematically as:

$$\langle \boldsymbol{p}^a, \boldsymbol{p}^t \rangle = \mathbf{A}\langle \boldsymbol{n}^a, \boldsymbol{n}^t \rangle \tag{3.3}$$

In this equation, $\boldsymbol{n}^a$ represents the audio representation of a noun, while $\boldsymbol{n}^t$ denotes its textual counterpart. The notation $\langle \boldsymbol{n}^a, \boldsymbol{n}^t \rangle$ indicates the concatenation of these two representations into a single composite vector, which captures the combined semantic information of the noun in both modalities. Similarly, $\boldsymbol{p}^a$ and $\boldsymbol{p}^t$ represent the predicted adjective-noun vectors corresponding to the audio and textual modalities, respectively, and their concatenation is represented as $\langle \boldsymbol{p}^a, \boldsymbol{p}^t \rangle$.

AT-Joint Regression The AT-Joint Regression method introduces a variant of the original regression formula in Equation 3.2, and is expressed as:

$$\boldsymbol{p}^a = \mathbf{A}\boldsymbol{n}^t$$

In this formulation, $\boldsymbol{p}^a$ represents the audio adjective-noun phrase vector, while $\boldsymbol{n}^t$ is the textual representation of the corresponding noun. This approach is distinct in that it utilises the textual noun representation as a guiding signal to train the adjective matrix $\mathbf{A}$.

AT-Concat Tensor Skipgram The AT-Concat Tensor Skipgram method is based on the modified training objective of the single-modality Tensor Skipgram (TSG) (Equation 3.1) and has the following objective function:

$$\sum_{(\boldsymbol{c'}^a,\boldsymbol{c'}^t)\in\mathcal{C}^a\times\mathcal{C}^t} \log\sigma\left(\mathbf{A}\langle\boldsymbol{n}^a,\boldsymbol{n}^t\rangle\cdot\langle\boldsymbol{c'}^a,\boldsymbol{c'}^t\rangle\right) + \sum_{(\overline{\boldsymbol{c}}'^a,\overline{\boldsymbol{c}}'^t)\in\overline{\mathcal{C}}^a\times\overline{\mathcal{C}}^t} \log\sigma\left(-\mathbf{A}\langle\boldsymbol{n}^a,\boldsymbol{n}^t\rangle\cdot\langle\overline{\boldsymbol{c}}'^a,\overline{\boldsymbol{c}}'^t\rangle\right) \tag{3.4}$$

Here, $\langle\boldsymbol{n}^a,\boldsymbol{n}^t\rangle$ represents the concatenation of the fixed pre-trained audio and textual embeddings of a noun, and $\mathcal{C}^a$ and $\mathcal{C}^t$ are the sets of positive and negative contexts of the adjective-noun phrase. For positive contexts, we utilise the fixed pretrained embeddings of the actual audio and text representations of the adjective-noun phrases. Conversely, for negative contexts, we fix the adjective and randomly choose a subset of nouns different from *n*. For example, consider learning the matrix $\mathbf{A}$ for the adjective *happy*. In this case, $\boldsymbol{n}^t$ is the textual embedding of *cat*, and $\boldsymbol{n}^a$ is the average of all its audio vectors. The term $\boldsymbol{c'}^a$ indexes over all the audio embeddings we have for *happy cat*, while $\boldsymbol{c'}^t$ is its textual embedding. For the negative contexts, $\overline{\boldsymbol{c}}'^a$ indexes over all the audio embeddings we have for *happy noun*, where *noun* is a randomly selected noun different from *cat*, such as *baby* or *car*.

AT-Joint Tensor Skipgram This method changes the objective function to the following, for the same $\boldsymbol{n}^t$ and $\mathcal{C}^a$ as above.

$$\sum_{\boldsymbol{c'}^a\in\mathcal{C}^a} \log\sigma\left(\mathbf{A}\boldsymbol{n}^t\cdot\boldsymbol{c'}^a\right) + \sum_{\overline{\boldsymbol{c}}'^a\in\overline{\mathcal{C}}} \log\sigma\left(-\mathbf{A}\boldsymbol{n}^t\cdot\overline{\boldsymbol{c}}'^a\right) \tag{3.5}$$

Here, the audio adjective is learnt from an audio-only context, but in such a way that when multiplied with the textual vector of a noun, it is forced to be closer to the audio context.

Audio-Only Models We additionally explored regression and skip-gram audio-only composition models. In the case of regression, we employ pre-trained auditory representations of adjective-nouns and nouns from OpenL3. Here, nouns are treated as independent variables, while adjective-noun-vectors function as dependent entities. On the other hand, the audio skip-gram model for phrase composition is defined by the following equation:

$$\sum_{\boldsymbol{c}_a' \in C_a} \log \sigma \left(\mathbf{A}\boldsymbol{n}_a \cdot \boldsymbol{c}_a' \right) + \sum_{\overline{\boldsymbol{c}_a}' \in \overline{C_a}} \log \sigma \left(-\mathbf{A}\boldsymbol{n}_a \cdot \overline{\boldsymbol{c}_a}' \right) \tag{3.6}$$

Where $\boldsymbol{n}_a$ is the pretrained aggregated audio noun embeddings, while C_a and $\overline{C}_a$ are the set of positive and negative contexts from pretrained audio representations.

Addition For two given vectors, one representing an adjective $\boldsymbol{A}$ and the other representing a noun $\boldsymbol{n}$, the additive composition of the adjective-noun pair (denoted as $\boldsymbol{p}$) is expressed as:

$$\boldsymbol{p} = \boldsymbol{A} + \boldsymbol{n} \tag{3.7}$$

We utilised Equation 3.7 to derive the auditory representations of adjective embeddings by estimating the adjective vector through the subtraction of the noun vector from the adjective-noun composition.

$$\boldsymbol{A}_{\text{est}} = \boldsymbol{p} - \boldsymbol{n} \tag{3.8}$$

$$\boldsymbol{p}_{\text{est}} = \boldsymbol{A}_{\text{est}} + \boldsymbol{n} \tag{3.9}$$

Subsequently, the estimated adjective-noun vector $\boldsymbol{p}_{\text{est}}$ is reconstructed by adding the noun vector to the estimated adjective vector. The subtracted nouns correspond to the averaged noun embeddings for each noun, while the added nouns are their non-averaged embeddings. The rest of the models can be found in Table 3.1[2]. In this table, the **ADD-Audio** follows a similar approach in which $\boldsymbol{a}^{an}$ corresponds to the averaged audio

[2]In earlier stages of model selection, we also explored some other alternatives, for instance, 1) multiplicative composition and 2) averaging all adjective-noun vectors to represent adjectives both in additive and multiplicative compositions. However, these methods did not demonstrate superior performance compared to the addition-subtraction approach. Moreoever, we implemented 3) a pilot study on text-only composition model, the details of which are provided in Appendix A.

embedding of the adjective-noun phrase, $\boldsymbol{n}^a$ refers to the averaged audio embedding of the noun across all its occurrences, and $\boldsymbol{n}'^a$ means the individual noun embedding used to reconstruct the phrase representation. More details in the next section.

Table 3.1: Phrase learning models and their formulations.

Model	Abbreviation	Formula
Non-Compositional Text	Non-Comp Text	$\boldsymbol{an}^t$
Non-Compositional Audio	Non-Comp Audio	$\boldsymbol{an}^a$
Additive Text	ADD-Text	$\boldsymbol{a}^t + \boldsymbol{n}^t$
Additive Audio	ADD-Audio	$(\boldsymbol{an}^a - \boldsymbol{n}^a) + \boldsymbol{n}'^a$
Additive Concatenation	ADD-AT	$((\boldsymbol{an}^a - \boldsymbol{n}^a), \boldsymbol{a}^t) + (\boldsymbol{n}^a, \boldsymbol{n}^t)$
Audio-Only Linear Regression	Audio-Only LR	$\boldsymbol{A} \times \boldsymbol{n}^a$
Joint Linear Regression	AT-Joint LR	$\boldsymbol{A} \times \boldsymbol{n}^t$
Concatenated Linear Regression	AT-Concat LR	$\boldsymbol{A} \times (\boldsymbol{n}^a, \boldsymbol{n}^t)$
Audio-Only Skipgram	Audio-Only SG	$(\boldsymbol{A} \times \boldsymbol{n}^a) \cdot \boldsymbol{c}'^a$
Joint Skipgram	AT-Joint SG	$(\boldsymbol{A} \times \boldsymbol{n}^t) \cdot \boldsymbol{c}'^a$
Concatenated Skipgram	AT-Concat SG	$(\boldsymbol{A} \times (\boldsymbol{n}^a, \boldsymbol{n}^t)) \cdot (\boldsymbol{c}'^a, \boldsymbol{c}'^t)$

3.3 Implementation

An overview of the methodology is presented in Figure 3.1, where (1) illustrates the audio-textual concatenation process, while (2) and (3) depict the single-modal and joint-learning approaches, respectively. In (2), the embeddings are utilised independently, whereas in (3), they are jointly optimised using matrix-based compositional models.

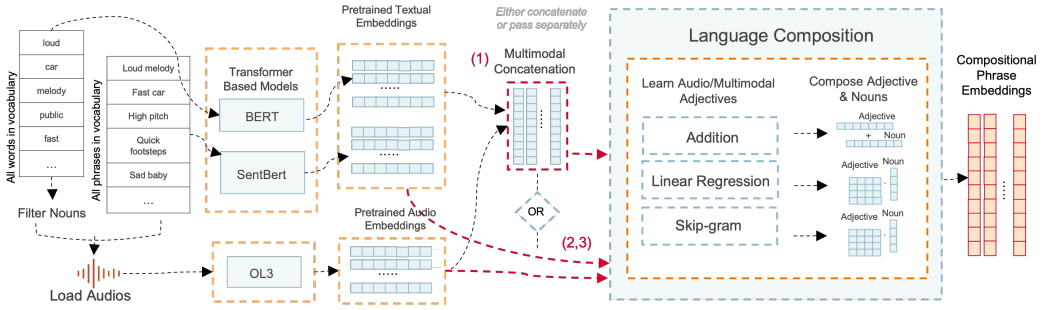


Figure 3.1: Overview of the methodology for combining the audio with the text

This section primarily details the implementation of the audio-text tensor skipgram model, inspired from the image-text composition model proposed by Lewis et al. [37]. The model's architecture and training procedure are explained below.

Dataset: The *Core* dataset constructed in Chapter 4 was used for training, and the *Sample* dataset was used for evaluation. The datasets consist of audio recordings associated with adjective-noun phrases and nouns. For each phrase and its corresponding noun, multiple audio samples were collected. For example, *loud cat* is represented by 82 audio files, while *cat* is represented by 100. For more details, refer to Chapter 4.

Vocabulary: For each adjective-specific model, all available audio embeddings are utilised across different adjective-noun phrases and nouns. Each model constructs an *expanded vocabulary* specific to the adjective, capturing the distinct auditory characteristics associated with each phrase. The size of the vocabulary varies for each adjective based on the number of nouns it modifies and the number of audio files available for each adjective-noun phrase. For example, adjective *fast* modifies 194 different nouns, and each adjective-noun pair is represented by 10 to 100 audio files. This leads to a comprehensive vocabulary of *6,657* distinct entries for adjective *fast*, where each entry corresponds to an audio embedding of a specific adjective-noun phrase modified by the adjective *fast*. For all adjectives, the vocabulary size varies from 2,429 to 8,478. This vocabulary is then used to construct positive and negative samples for training the skip-gram model[3].

Context: The **positive context** consists of multiple audio files representing the target adjective-noun phrase. For instance, *loud melody* is represented by 100 audio files, while *loud cat* is represented by 82 files. These audio files are treated as the positive context for each phrase, as they all correspond to the same target phrase. Using all available examples for a given phrase as positive context helps the model recognise it across diverse auditory conditions, such as variations in background noise, speaker intonation, or context.

In contrast, **negative samples** are drawn from unrelated noun phrases that share the same adjective but differ in meaning (e.g., *loud car* or *loud bell*). This approach increases the difficulty of distinguishing between correct and incorrect contexts, compelling the model to learn the subtle distinctions in how different nouns are modified by the same

[3]Although the vocabulary size for some adjectives is relatively small, the rich averaged noun vectors associated with each adjective ensure adequate representation for computing compositions. These embeddings are robust, as they are derived from averaging a substantial number of sound files for each noun corresponding to the adjective. Notably, the number of audio representations for nouns per adjective may go from 2,778 to 20,894 sound files.

adjective. To prevent negative samples from being too similar to the positive context (e.g., *loud music* and *loud melody*), the selection of negative samples is kept dynamic. For this, we treated the selection of negative nouns as a hyper-parameter, refining the choice through tuning on the validation set. Specifically, we randomly generate 10 different sets of negative samples, run 50 epochs for each set, and select the best-performing set. This process ensures that negative samples progressively become harder as the model improves, enhancing its overall learning.

Example: For each positive example (target word and positive context), one corresponding negative example is randomly sampled. The format is:

$$((\text{pos\_phrase}, \text{target}), \text{neg\_phrase})$$

For a given target phrase, for instance *fast_car*, the training samples might look like:

$$((\textit{fast\_car1}, \textit{fast\_car}), \textit{fast\_music1}), ((\textit{fast\_car2}, \textit{fast\_car}), \textit{fast\_steps20}), \ldots, ((\textit{fast\_car100}, \textit{fast\_car}), \textit{fast\_bus50})$$

Each tuple represents a training example for the skip-gram model with negative sampling. Finally, the dataset is split into train (80%), validation (10%), and test (10%) sets. Care is taken to ensure that no overlapping positive classes exist between these splits, ensuring that the model does not encounter the same positive context during both training and testing.

Training: For skipgram models, the learning rate was 10^{-6} with a batch size of 512, and a training duration of 200 epochs. The models were trained on NVIDIA T4 and V100 on Google Colab. The training was done in batches over a period of 3 months, totalling ~100 hrs. We used Binary Cross-Entropy loss and the Adam optimiser in the training process to refine the performance. Principal Component Analysis (PCA) was used to equalise the dimensions of auditory and textual representations to 50.

Algorithm 1 details the training procedure for the **AT-Concat** tensor Skipgram model, processing adjective-noun pairs from the dataset *S*. For each pair, audio and text embeddings are concatenated into a multimodal representation, with negative distractors randomly sampled from the noun set, excluding the target noun. The training involves calculating positive and negative scores based on the concatenated embeddings and estimating the probability of the target phrase using a softmax function, with loss computed via cross-entropy and weight decay for regularization. After each epoch, validation

Algorithm 1 Algorithm to train adjective-noun compositions via AT-Concat

Input: Training dataset S, audio encoder $\mathcal{E}_a$, text encoder $\mathcal{E}_t$, composition encoder $\mathcal{C}$, learnable parameters θ, adjectives $\mathcal{A}$, nouns $\mathcal{N}$, weight decay λ, number of epochs M, validation accuracy thresholds $\mathcal{T}$, validation check interval k.
Output: Learned model parameters θ.
for $i \leftarrow 1$ to M **do**
    for all $(x,y) = (a,n) \in S$ **do**
        $x_{\text{audio}}, x_{\text{text}} \leftarrow \mathcal{E}_a(x), \mathcal{E}_t(x)$ ▷ Get embeddings for the positive context
        $x \leftarrow [x_{\text{audio}}; x_{\text{text}}]$ ▷ Concatenate audio and text embeddings
        Sample negative distractor y_{neg} from $\mathcal{N} \setminus \{y\}$
        $l_{\text{pos}} \leftarrow x \cdot \mathcal{C}(x,y)$
        $l_{\text{neg}} \leftarrow x \cdot \mathcal{C}(x, y_{\text{neg}})$
        $p_\theta(y = (a,n)|x) \leftarrow \frac{\exp(l_{\text{pos}})}{\exp(l_{\text{pos}}) + \exp(l_{\text{neg}})}$
        $L \leftarrow -\log p_\theta(y|x) + \lambda \|\theta\|_2$ ▷ Cross-entropy loss with weight decay
        Update θ
        Compute the validation accuracy A_{val}
        if i equals k **then**
            if $A_{\text{val}} \geq \tau$ **then** ▷ Check the accuracy threshold
                Continue training for $(M-k)$
            else
                Select new y_{neg} and repeat from step 2
            end if
        end if
    end for
    if no set of negative distractors reaches $A_{\text{val}} \geq \tau$ after all trials **then**
        Repeat from step 7.
    end if
end for
return θ

accuracy A_{val} is evaluated based on human judgments of phrase similarity, with checks performed every k epochs—a value ranging from 20 to 30 depending on the dataset size and computational resources. If A_{val} meets or exceeds the highest threshold τ (values from 0.7, 0.6, 0.5), training continues with the current distractors. If not, the threshold is systematically reduced to the next value, and this process is repeated until reaching the lowest threshold. If the lowest threshold is reached without satisfactory performance, a new trial is initiated, and a new set of distractors is selected, with a maximum of three trials allowed. Should all trials be exhausted without meeting the threshold, the model retains the last selected distractors to ensure training progresses.

Linear Regression: In this method, corpus-based nouns act as independent variables, and adjective-noun vectors serve as dependent variables. Adjective-noun embeddings

are scaled for uniformity, and the dataset is split with 80% for training and 20% for testing. Adjectives are modeled as linear functions using Partial Least Squares Regression (PLSR) [110], with a coefficient matrix computed from averaged noun and adjective-noun embeddings (Figure 3.1). The number of latent components for each model was optimised to enhance performance. The resultant matrix is then multiplied by corpus-based non-averaged noun vectors to generate multiple phrase representations, which are averaged into a single representation per phrase.

Addition: The process begins by reading all corpus-based adjective-noun embeddings. For each phrase, which may possess multiple embeddings, the mean embedding is calculated to derive a representative vector. Following this, all corpus-based noun embeddings are retrieved, and the mean embedding for each noun is computed accordingly. The adjective vector is then derived by subtracting the averaged noun embedding from the averaged phrase embedding (Figure 3.1). This process generates multiple vectors for each adjective, based upon the number of nouns it modifies. For instance, the adjective *fast* modifies 194 distinct nouns, resulting in 194 unique vectors associated with this adjective. These vectors are subsequently averaged to yield a singular, comprehensive adjective representation.

3.4 Conclusion

This chapter introduces a novel multimodal compositional distributional semantics framework, integrating audio features with linguistic data. It presents a formalism and methods for grounding and composing adjective-noun phrases, while effectively capturing their semantic and auditory interactions. Adjectives are represented as matrices and nouns as vectors, extending the compositional distributional semantics framework to the audio-text domain. The next chapter introduces a dataset designed to evaluate the effectiveness of the proposed framework in capturing phrase similarities.

Chapter 4

A Novel Multimodal Phrase Dataset

This chapter addresses the limitations of existing phrase similarity benchmarks by introducing a novel multimodal dataset that captures both semantic and auditory similarities between adjective-noun phrases.

Over the last decade, there has been significant emphasis on how the language can be grounded in vision, with numerous studies integrating images to enhance semantic understanding [27–30, 37, 38]. However, the auditory modality has remained relatively underexplored, with only a few researchers (e.g., Lopopolo & Miltenburg [32] and Kiela & Clark [33]), exploring the integration of audio into semantic models. This limited focus is primarily due to the scarcity of comprehensive training and evaluation datasets. Existing similarity benchmarks primarily focus on textual analysis [4, 10, 29] and often lack adjective pairs relevant to auditory contexts. Word similarity datasets like MEN [11] and WS353 [10] are limited to abstract or cultural concepts, while AMEN [33] addresses sound relevance only at the word level. Phrase similarity benchmarks, particularly for adjective-nouns, such as the one proposed by Mitchell and Lapata [41], face similar limitations. Chapter 2 discusses these datasets and their limitations in detail.

To address this gap, this chapter presents a dataset that captures semantic relationships between adjective-noun phrases in text and their meaningful auditory associations. The proposed dataset is divided into two subsets: **SemPhrase** for semantic similarities and **AudPhrase** for audio similarities. The chapter offers a comprehensive overview of the dataset construction, starting with the methods employed to identify and select sound-relevant adjectives and nouns. It then explores the annotation process, detailing the

criteria and guidelines provided to human annotators and concludes by discussing the inter-annotator agreement results.

4.1 Building the Dataset

This section explains the detailed method used to create a textual-auditory dataset, made especially for adjective-noun phrase composition. The process includes selecting sound-relevant adjectives, filtering and pairing nouns, and finally building and validating the dataset. Figure 4.1, shows a visual summary of the systematic approach we followed, highlighting the main stages of data collection and refinement. Each of these stages is explained in detail in the following sections.

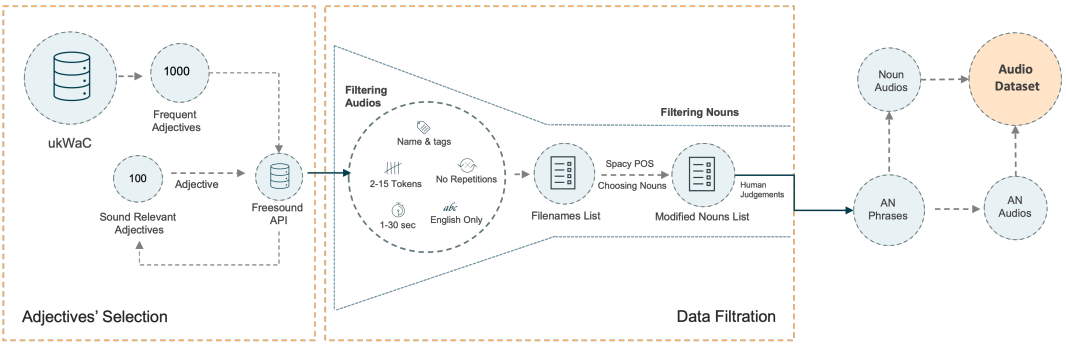


Figure 4.1: Multimodal phrase data construction

4.1.1 Selecting Adjectives

The first step in building the dataset was to carefully choose adjectives that are commonly used in the English language and are also relevant in an auditory context[1]. This included:

1. Textual Relevance: The UKWaC [111] is used to extract 1,000 most frequently used adjectives in English. UKWaC is a large-scale, web-derived collection of English text, containing over 2 billion words. It was created by crawling the .uk domain to gather a wide range of text types, from academic articles to blogs, and is enriched with linguistic annotations like part-of-speech tagging and lemmatization. This makes it a highly valuable resource for English linguistic research.

[1] As previously discussed, in this thesis, **auditory relevance** is defined as: *An adjective, when combined with a noun, evokes or is associated with a specific sound*, as seen in phrases like *creaky door* or *loud horn*.

2. Acoustic Relevance: Once the frequent English adjectives were identified, the next step was to assess their relevance in an acoustic context. This was achieved by cross-referencing each adjective with the Freesound library[2], an extensive online repository of sound samples. The aim was to see if these adjectives were commonly paired with nouns in settings that explicitly involve sound. An adjective was deemed acoustically relevant if it appeared with a noun in more than 800 instances within the Freesound database, based on the names or tags of the audio files.

3. Multimodal Adjectives: Building on the previous steps, a final set of 100 adjectives was carefully selected to ensure both linguistic and acoustic significance. This set was further narrowed down to 30 adjectives (given in Appendix B.1) that exhibited strong acoustic associations. The refinement process prioritised adjectives that were paired with at least 25 unique nouns, each represented by a minimum of 100 sound files in the Freesound database.

4.1.2 Selecting Sounds

After selecting the adjectives, the next step was to gather all the sounds associated with each adjective. The following criteria was applied to collect and filter sounds:

(a) **Name or Tag:** For each adjective (e.g., *Fast*), only files from Freesound where the adjective appeared in either the *Name* or *Tag* were selected.
(b) **Filename Length:** Filenames containing 2 to 15 tokens were considered, excluding overly simplistic filenames like *fast.ogg*.
(c) **Duration:** The audio files needed to be between 1 and 30 seconds long to ensure relevance without being too brief or extended.
(d) **Repetitions**: Repeated filenames were excluded to maintain a diverse set of examples.
(e) **Names:** Only English filenames were included to ensure consistency in language.
(f) **Adjective-Noun Pairing:** Filenames include at least one adjective and noun, such as *fast shutter* or *fast walk*, ensuring contextually relevance in phrases.

[2]https://freesound.org

4.1.3 Selecting Nouns

After identifying the adjectives and gathering associated sound files, the next step is to filter nouns to pair with these adjectives. This involves:

1. Identifying Nouns: After preparing the list of audio filenames, Spacy POS tagging was employed to identify nouns within each filename. As filenames often contain multiple nouns, the first identified noun is selected as the one modified by the adjective. This is because the automated methods like dependency parsing were deemed unsuitable due to the frequent occurrence of poor grammar, nonsensical terms, and phrases lacking adjectives in filenames (e.g., instances where the adjective is present only in the tags). Table 4.1 illustrates examples.

Table 4.1: Filenames from FreeSound where the selected nouns were not always meaningful

Adjective	Filename	Nouns	Selected Noun
loud	CS 80 PWM FAST - 78 (F#5) - vel 127	['pwm', 'f', 'vel']	pwm
distant	auto distant heard from construction site	['auto', 'construction', 'site']	auto
fast	64x fast-forward speech effect (spooling)	['64x', 'speech', 'effect']	64x

2. Filtering Meaningful Nouns: To ensure the selected nouns were meaningful, a filtering process was applied to the filenames. First, only filenames containing English nouns were considered. Next, plural nouns were converted to their singular form to eliminate redundancies. Finally, meaningless nouns were removed using Python-based spellchecker. Table 4.2 illustrates examples, where for *path_next_to_rail*, the adjective *fast* was found in its tags.

Table 4.2: Filtered nouns and resulting adjective-noun (AN) phrases.

Adjective	Filename	Nouns	AN Phrase
quick	Shutter Camera, quick, A	['shutter', 'camera']	quick shutter
fast	path_next_to_rail	['rail']	fast rail
loud	Loud Traffic on the Highway	['traffic', 'highway']	loud traffic

3. Manual Review: Even after filtering, some phrases lacked meaningful context. For example, while *fast file* contains a relevant sound, the word *file* is ambiguous in its meaning. To ensure the quality of these pairs, a manual review was conducted, following Kiela & Clark [33]. Each adjective-noun pair was carefully evaluated by authors to retain

only those with clear meanings. Nouns that were ambiguous or unclear, either in text or audio, were removed from the final selection. Table 4.3 presents a snippet of this review process. Some other discarded nouns include *bat, bank, jam, hip, tape, rail* etc.

Table 4.3: Manual review of adjective-noun phrases.

Adjective	Selected Noun	Relevance
quick	shutter	Y
fast	file	-
fast	rail	Y
loud	traffic	Y

4. Additional Data: After finalizing the list of meaningful pairs of noun phrases, additional data was collected from Freesound for each pair. For the selected nouns, 100 audio files were gathered for each; and 10-100 for each adjective-noun combination due to varying usage frequencies, all in the standard open-source OGG format.

4.1.4 The Dataset

The final dataset was divided into two main parts: *The Core* and *The Sample*. The *Core* dataset serves as a comprehensive resource for large-scale training, offering a broad range of data, while the *Subset* dataset provides a more focused collection specifically designed for gathering human judgments for final evaluations.

4.1.4.1 The Core

This dataset consists of **30** adjectives, **1,944** nouns, and **92,157** pairs of noun phrases. In the dataset, the number of nouns modified by each adjective varies; e.g. *low* modified 46 nouns, while *quick* modified 114, with an average of 65 nouns per adjective. For audios, we selected 100 audio files per noun and on average 50 files per adjective-noun. The number of audio files per adjective-noun varied, e.g., 97 for *human cough* and 45 for *angry girl*. In total, the dataset contained **271,766** audio files, equivalent to approximately **760** hours of audio data.

Example phrases from the dataset: *big drum, dark music, angry grunt, loud thunder, distant blast, digital beep, big laughter, melodious voice, sad dog, heavy punch, happy child, fast typing, female robot, angry monster, and sad music.*

4.1.4.2 The Sample

For human judgements, we used a subset of the Core dataset for which we excluded adjective-noun combinations with fewer than 50 associated audio files. This resulted in a reduced number of combinations per adjective, ranging from 15 to 20 nouns per adjective. This decision was made to strike a balance in the dataset, ensuring that all adjectives had a roughly equal number of nouns and sounds, thereby facilitating a fair and unbiased evaluation. As a result, the **sample dataset** consists of **30** adjectives, **524** nouns, and a total of **2,950** pair of noun phrases. For audios, each noun is represented by **100** sound files, while each adjective-noun has a range of 30 to 100 sound files, and **96,794** sound files in total.

4.2 Human Annotations

To assess the quality of the learned adjective-noun representations, human judgments were collected for the *sample* dataset with two primary objectives in mind: The **first objective** was to evaluate the **semantic similarity** between pairs of noun phrases. Annotators were asked to consider the degree of similarity in semantic meaning between different noun phrase pairs. For example, they might assess how semantically similar the phrases *loud piano* and *loud music* are. The **second objective** was to evaluate the **sound similarity** of pairs of noun phrases. This task required annotators to imagine and interpret the auditory qualities associated with each phrase and then judge how similar these auditory characteristics might be. For instance, they might evaluate how similar the phrases *loud horn* and *soft horn* are in terms of the sound they evoke.

4.2.1 Categorization

Pilot Study: To explore how people perceive and evaluate both semantic and auditory similarities, a pilot study was conducted. In this study, 10 annotators were tasked with evaluating 100 random phrase pairs drawn from 6 different adjectives in the dataset, across both semantic and auditory similarity dimensions. It yielded an inter-annotator agreement of 0.45 while highlighting several challenges, particularly related to the clarity of noun meanings, which can vary significantly depending on context. For instance, the word *slap* might be understood differently in an environmental context (e.g., the

sound of a hand slapping) compared to a musical context (e.g., a sharp musical hit). These contextual ambiguities posed challenges for the annotators, leading to potential confusion in evaluations.

Classification: To address this challenge, a strategy was implemented to categorise phrases based on the nouns associated with each adjective into either environmental or musical contexts. For example, the phrase *fast car* was classified as environmental, while *fast drum* was categorised as musical. During the annotation process, this classification ensured that phrases were only compared with others from the same context, environmental phrases were paired exclusively with other environmental phrases, and musical phrases with other musical phrases. This strategy helped maintain consistency and relevance in the similarity judgments provided by the annotators, allowing for more accurate within-category comparisons.

In the sample dataset, there were **2,392** pairs categorised as **environmental** and **558** pairs as **musical**. To facilitate the annotation process and enhance the annotators' understanding of the meaning or sound of each phrase, environmental and musical phrases were presented separately for each task. For further details, see Appendix B.1.

> Some examples of environmental pairs are: *(angry grunt, angry girl), (distant car, distant blast), (big laughter, big guitar)* and musical pairs are: *(melodic drum, melodic beat), (musical flute, musical guitar), (sad music, sad guitar)*.

4.2.2 Elicitation Procedure

The elicitation procedure for annotations involved conducting online studies via Amazon Mechanical Turk[3]. Each adjective-noun pair was rated by **15** subjects, either for semantic or sound similarity, using a scale ranging from 1 to 5, with 1 indicating the lowest level of similarity (For example, see Figure 4.2). Only individuals from English-speaking countries with a HIT approval rate above **95%** and more than **1,000** approved HITs were allowed to participate. The pairs of noun phrases were divided into batches, and each batch included two trick questions designed as quality checks. These trick questions involved pairs with identical phrases, aiming to identify potential automated or inattentive responses. Additionally, the time taken by each annotator to complete

[3]https://www.mturk.com

the task was recorded; annotations completed significantly faster than the expected time were excluded from the dataset to prevent low-quality contributions. Through this procedure, **44,250** annotations were collected for each task (semantic and auditory similarity), resulting in a total of **88,500** annotations.

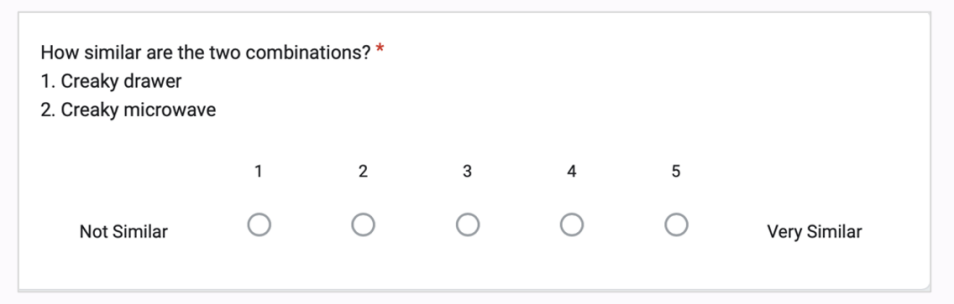


Figure 4.2: An example question for annotation.

4.2.3 Results: Inter-Annotator Agreements

After collecting annotations, we measured inter-annotator agreement for each adjective–noun pair.

Evaluation Method: Inter-annotator agreement was measured using a standard method proposed by Hill et al. [9], which compares each annotator's ratings with the average ratings of all other annotators on the same items (leave-one-out). This approach treats the mean of other raters as the reference standard, allowing for consistent evaluation even when annotators rate different subsets of data. It is robust to missing annotations and avoids distortions caused by computing correlations over disjoint sets. Spearman's rho was used to calculate the correlation for each annotator, and the final agreement score is reported as the average across all annotators. The results are shown in Table 4.4.

Table 4.4: Inter-annotator agreement scores for semantic and audio similarity tasks

Average correlations	**Semantic Similarity**		**Audio Similarity**	
	Env.	**Mus.**	**Env.**	**Mus.**
Per batch				
Min	0.69	0.66	0.66	0.62
Max	0.72	0.7	0.7	0.71
All batches	0.7	0.69	0.69	0.65
Overall	0.69		0.67	

In the semantic similarity task, the average correlation per batch ranged from 0.69 to 0.72 for environmental pairs and from 0.66 to 0.70 for musical pairs, with overall averages of 0.70 and 0.69, respectively. In the audio similarity task, the average correlation per batch ranged from 0.66 to 0.70 for environmental pairs and from 0.62 to 0.71 for musical pairs, with overall averages of 0.69 and 0.65, respectively. Combining both environmental and musical pairs, the overall average correlation was 0.69 for semantic similarity and 0.65 for audio similarity.

What causes disagreements? Disagreements arise in the cases where annotators differ in their understanding or perception of phrases. For example, for semantic similarity of the pair *(fast runner, fast internet)*, a significant disparity stems from interpreting *fast* in the context of physical versus abstract entities. Similarly, for the audio similarity between *(heavy rain, heavy noise)*, variations in perception may stem from differences in how annotators relate *heavy* to the auditory characteristics of natural versus artificial sounds.

The results indicate a high level of agreement and consistency among the human annotations across both the semantic similarity and audio similarity tasks, underscoring the reliability of the data collected. For simplicity, we will refer to the annotated semantic similarity dataset as **SemPhrase** and the audio similarity dataset as **AudPhrase** for the rest of this document. These annotations are made publicly available on Github[4].

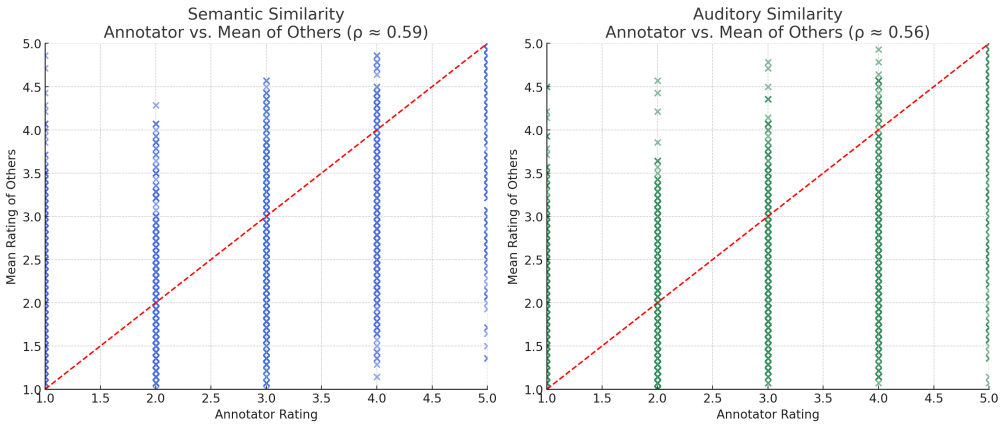


Figure 4.3: Scatter plots comparing one annotator's ratings with the mean of the remaining annotators. Left: Semantic similarity ($\rho \approx 0.59$). Right: Auditory similarity ($\rho \approx 0.56$).

To illustrate inter-rater consistency, Figure 4.3 shows scatter plots for an annotator whose Spearman's ρ is closest to the overall average. Each point represents a phrase pair,

[4]https://github.com/audio-comp

plotted by its score from the selected annotator (x-axis) and the mean score from all other annotators (y-axis). The vertical alignment of points reflects the use of discrete integer ratings from 1 to 5 by the selected annotator, while the y-axis shows the average of other annotators. The dashed red diagonal indicates perfect agreement. The plots include both environmental and musical phrase pairs, providing a combined view across both subcategories. While many points fall near the diagonal, there is noticeable variation in the group means for each rating level, especially in the auditory task. This spread reflects the subjective nature of the task and suggests that, although there is general agreement, annotators often differ in the exact similarity scores they assign. The overall pattern still indicates a positive trend and supports the consistency and quality of the collected annotations. For further details, see Appendix B.1.

4.3 Conclusion

This chapter addresses the limitations of existing datasets in evaluating multimodal adjective-noun phrase similarities by introducing a novel textual-auditory dataset. Sound-relevant adjectives and nouns were systematically selected to ensure their relevance to both linguistic and auditory contexts. A rigorous filtering and validation process, including human annotations, refined the dataset and ensured its quality. The development of a two-part dataset structure is outlined: a comprehensive core dataset for large-scale training and a refined sample dataset for human evaluations. The collected annotations and their analysis validate the dataset's reliability and highlight its potential for multimodal phrase similarity tasks. The next chapter leverages this dataset to evaluate semantic and audio phrase similarity tasks, demonstrating its effectiveness in capturing the relationship between text and sound.

Chapter 5

Evaluation of the Framework

This chapter aims to investigate the effectiveness of multimodal compositional models (MultiCoDi) in capturing both semantic and auditory similarities between adjectives and adjective-noun phrase pairs.

This chapter provides a comprehensive analysis of how the integration of textual and auditory data enhances phrase-level understanding, while addressing several key questions. First, it asks: *Can combining text and audio data in a multimodal setting (such as through concatenation and joint learning) outperform models relying solely on audio?* To investigate this, audio-only variants of regression and tensor skip-gram models were trained, learning adjective matrices from audio vectors tied to their corresponding nouns and contexts. The chapter also examines whether *non-commutative models, such as regression and tensor skip-gram, outperform simpler commutative models*. To test this, an additive model was implemented, combining the representations of adjectives and nouns. Lastly, the performance of *compositional models* is compared to *non-compositional approaches* by evaluating them against holistic OpenL3 audio vectors of adjective-noun phrases, assessing both semantic and auditory relationships.

The chapter is structured as follows: First, the evaluation methods are outlined, including the use of cosine similarities, Spearman correlations, and matrix similarities. Next, the results of the analysis on adjective similarities are presented, comparing different models across various techniques. Phrase similarities are then explored, evaluating the models' performance on both semantic and audio similarity tasks. Next, the analysis section investigates the learned adjective-noun embeddings through K-means clustering. Finally, the chapter concludes with a summary of key findings.

5.1 Evaluation Methods

Various techniques were employed to evaluate the performance of the models. This section details the evaluation methods applied across all tasks.

5.1.1 Cosine Similarities

Cosine similarity measures the cosine of the angle between two non-zero vectors in an inner product space, providing a similarity score between -1 and 1 and is given by the following equation:

$$\text{cosine similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{5.1}$$

Where:

- $\mathbf{A}$ and $\mathbf{B}$ are the vectors being compared.
- $\mathbf{A} \cdot \mathbf{B}$ is the dot product of the two vectors.
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the magnitudes (or lengths) of the vectors.

A cosine similarity of 1 indicates perfect similarity, meaning the vectors point in the same direction. A value of 0 indicates no similarity between the vectors, as they are orthogonal and share no common direction or relationship in the vector space. In contrast, a value of -1 represents perfect dissimilarity, where the vectors point in completely opposite directions, indicating maximum divergence in meaning or context. This negative correlation typically occurs when the concepts being compared are highly contradictory or unrelated within the given embedding space.

Cosine similarity is a widely utilised metric for comparing vectors that represent words, sentences, or documents in a high-dimensional space. These similarity scores are often evaluated against human similarity judgments by calculating their correlation. A high correlation between model-generated scores and human judgments suggests that the embeddings effectively capture semantic relationships, aligning closely with human understanding of linguistic meaning.

5.1.2 Spearman Correlations

Spearman's rank correlation coefficient (ρ) is a non-parametric statistical measure used to evaluate the strength and direction of the monotonic relationship between two ranked variables. It is calculated using the following formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{5.2}$$

Where:

ρ is Spearman's rank correlation coefficient.

d_i is the difference between the ranks of corresponding values of the two variables (e.g., model scores and human judgments).

n is the number of observations.

It starts by assigning ranks to the values in each dataset, with the smallest value receiving rank 1. Next, compute the differences (d_i) between the ranks of corresponding observations. Square these rank differences to obtain d_i^2, and sum all the squared differences to compute $\sum d_i^2$. Finally, use these values in the Spearman formula to calculate ρ. The value of ρ ranges from -1 to 1, indicating the strength and direction of the association between the ranked variables. A ρ of 1 signifies a perfect positive correlation, where higher ranks in one variable correspond exactly to higher ranks in the other. Conversely, a ρ of -1 indicates a perfect negative correlation, where higher ranks in one variable correspond exactly to lower ranks in the other. A ρ of 0 indicates no correlation between the variables.

To evaluate how well the model's similarity measures align with human judgments, Spearman's coefficient (ρ) is used in this thesis. This method ensures that the relationship between the model's predictions and human judgments is effectively measured, even when the relationship is not strictly linear. It provides a reliable assessment of how well the models capture human intuition in both semantic and auditory similarity tasks.

5.1.3 Matrix Similarities

For adjective matrix calculations, we adopted the method proposed by Maillard and Clark [22], who demonstrated that cosine similarity is inadequate for comparing adjective *matrices* due to its poor correlation with human judgment standards. Instead, they suggest measuring adjective matrix similarity by examining how similarly they transform nouns[1].

[1]Maillard & Clark [22] argued that cosine similarity, though commonly used to compare vectors, does not adequately capture the functional role of matrices as linear transformations. They recommended evaluating adjective matrix similarity based on the extent to which matrices produce comparable transformations of noun vectors, aligning more closely with human judgments. This

This is done by computing the cosine distance between noun images under two adjectives, using k-means cluster centroids. For two adjectives, **A** and **B**, the similarity is calculated using the following formula:

$$\text{matsim}(\mathbf{A},\mathbf{B}) = \underset{\boldsymbol{n}\in\mathcal{N}}{\text{median}}\left(\text{vecsim}(\mathbf{A}\boldsymbol{n},\mathbf{B}\boldsymbol{n})\right) \tag{5.3}$$

Where $\boldsymbol{n}$ represents a vector centroid from a set of noun clusters $\mathcal{N}$, and the median is taken over all centroids. This approach allows for a more accurate comparison of how two adjectives, represented by matrices, modify nouns in semantic space.

For this thesis, to compute matrix similarities, our approach begins by calculating the cluster centroids for all holistic auditory phrases within the dataset using k-means clustering. We then apply the model-based matrices to the median of these centroids. Afterward, cosine similarities between phrase pairs are computed, and then compared to human judgments.

5.2 Adjective Similarities

5.2.1 Dataset

Semantic similarities between model-based adjective pairs are evaluated against the gold-standard Simlex-999 [9], a widely used benchmark for measuring word similarity. Simlex contains human-annotated similarity scores for each word pair, ranging from 0 (no similarity) to 10 (maximum similarity). Overlaps between adjectives in the audio dataset and Simlex were identified, with 11 of the 30 adjectives from the core dataset found in Simlex. This subset, referred to as Aud-SIMLEX, was used to compute pairwise similarities for all adjective pairs.

Of the 30 adjectives in the core dataset, 11 were found to be audio-relevant in Simlex. These adjectives are: *cheerful, rapid, happy, fast, large, huge, quick, angry, big, heavy,* and *young*.

approach emphasises the functional behavior of matrices over their geometric properties.

5.2.2 Evaluation Technique

For *non-compositional* adjectives methods, like ADD method, which yields adjective vectors, cosine similarities were computed between the vectors corresponding to pairs of adjective. Spearman correlations were then calculated by comparing these cosine similarities with the similarity scores provided by Simlex annotators for the respective pairs. For *compositional* methods, the matrix similarity computation discussed in Section 5.1.3 was applied.

5.2.3 Results

Table 5.1 reports the semantic similarity scores for adjective pairs, comparing three models (Audio-Only, AT-Joint, and AT-Concat) across three composition methods: Addition, Linear Regression, and Tensor Skip-Gram (TSG).

Table 5.1: Semantic similarities between adjectives.

Model	Simlex-Audio	
	Linear Regression	**Tensor Skipgram**
AT-Concat	0.73	0.76
AT-Joint	0.64	0.79
Audio-Only	0.68	0.74
ADD-Audio	0.46	
ADD-AT	0.50	
Non-Comp Text	0.65	

Across methods, adjective matrices generally achieve higher correlations than adjective vectors, consistent with their ability to represent more complex interactions between adjectives and their associated nouns. Among the composition methods, TSG consistently produces the highest values, with the strongest performance observed for **AT-Joint** (0.79) and **AT-Concat** (0.76).

When comparing multimodal models to unimodal baselines, the inclusion of both text and audio often results in stronger alignment with human semantic similarity judgements. For example, under Linear Regression, **AT-Concat** scores 0.73 compared to 0.68 for Audio-Only and 0.65 for Non-Comp Text. With TSG, **AT-Joint** reaches 0.79 and **AT-Concat** 0.76, both outperforming the Audio-Only (0.74) and Non-Comp Text (0.65)

baselines. These patterns suggest that combining modalities can capture complementary information that improves representation quality.

Overall, multimodal models paired with TSG tend to produce the strongest correlations, with consistent gains over both unimodal and additive approaches. A small-scale adjective-adjective evaluation of these results via bootstrapping is provided in Appendix C.

5.3 Phrase Similarities

5.3.1 Dataset

We evaluated the learned phrase embeddings using the SemPhrase and AudPhrase datasets, introduced in Chapter 4, section 4.2, to assess semantic and audio similarities between phrase pairs.

5.3.2 Evaluation Technique

To evaluate the models' ability to capture semantic and audio similarities, cosine similarities are first computed between the model-generated embeddings of adjective-noun pairs. These pairwise similarities are then compared to human annotations by calculating the Spearman correlation (ρ_s). The evaluation results for this experiment have been summarised in Table 5.2.

5.3.3 Results

The phrase similarity task results are summarised in Table 5.2. For semantic similarity (Table 5.2a), the non-compositional text baseline (TSG) achieves 0.71, while both AT-Joint and AT-Concat score higher with 0.88 and 0.86, respectively. For audio similarity (Table 5.2b), the non-compositional audio baseline records 0.58, with AT-Joint and AT-Concat reaching 0.89 and 0.88. These results show that compositional models outperform their respective unimodal baselines in both tasks.

When comparing non-commutative composition methods (Linear Regression and TSG) to commutative additive models, additive baselines generally achieve lower values. In the semantic task, additive baselines with TSG record 0.69 (ADD-Audio) and 0.65

Table 5.2: Models' performance in semantic and audio similarity tasks.

(a) Semantic Similarity

Model	Semantic Similarity	
	Linear Regression	**Tensor Skipgram**
AT-Concat	0.76	0.86
AT-Joint	0.67	0.88
Audio-Only	0.72	0.78
ADD-Audio	0.69	
ADD-AT	0.65	
Non-Comp Text	0.71	

(b) Audio Similarity

Model	Audio Similarity	
	Linear Regression	**Tensor Skipgram**
AT-Concat	0.78	0.88
AT-Joint	0.58	0.89
Audio-Only	0.75	0.83
ADD-Audio	0.74	
ADD-AT	0.67	
Non-Comp Audio	0.58	

(ADD-AT), compared to 0.88 for AT-Joint. In the audio task, additive baselines reach 0.74 (ADD-Audio) and 0.67 (ADD-AT), again below the 0.89 achieved by AT-Joint.

Multimodal models also tend to outperform unimodal alternatives. For semantic similarity, the highest score is 0.88 (AT-Joint, TSG), followed by 0.86 (AT-Concat, TSG) and 0.78 (Audio-Only, TSG). Under Linear Regression, AT-Concat achieves 0.76, AT-Joint 0.67, and Audio-Only 0.72. For audio similarity, the highest score is 0.89 (AT-Joint, TSG), with AT-Concat at 0.88 and Audio-Only at 0.83. Under Linear Regression, AT-Concat scores 0.78, AT-Joint 0.58, and Audio-Only 0.75.

Across both semantic and audio similarity tasks, TSG models consistently outperform all other approaches, whether commutative, non-commutative or non-compositional, with multimodal variants often achieving the highest correlations. Given this consistent advantage, we focus our bootstrapping analysis on TSG models to examine the stability of their performance in more detail.

5.3.4 Analysis

This section investigates how stable the top-performing TSG models are when faced with variability in the evaluation data.

Figure 5.1 shows the bootstrap distributions of Spearman's rank correlations between model-predicted similarities and human similarity ratings for the semantic (top row) and audio (bottom row) phrase similarity tasks. The left column shows results for the best-performing multimodal model (AT-Joint TSG), and the right column shows a weaker unimodal baseline (Non-Comp). These models were selected as representative cases, offering a focused illustration of correlation variability under resampling. The resulting distributions provide an indicative rather than exhaustive view of potential shifts in model-level correlations under resampling.

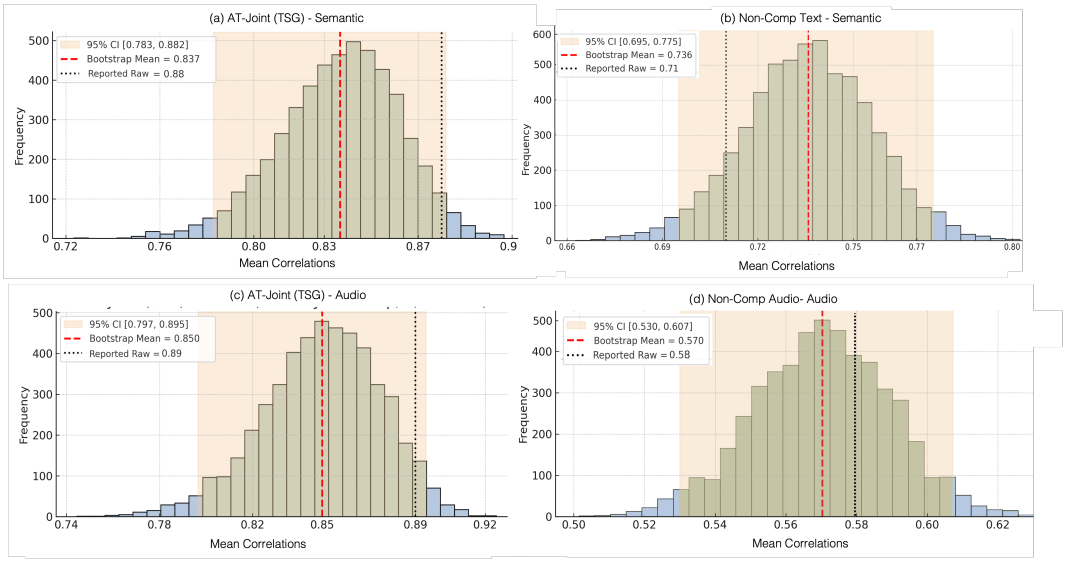


Figure 5.1: Bootstrap distributions of Spearman's rank correlations between model-predicted similarities and human ratings for the semantic (top row) and audio (bottom row) phrase similarity tasks. Left: best-performing multimodal model (AT-Joint TSG). Right: unimodal baseline (Non-Comp).

A two-way bootstrap was applied, resampling both adjectives and human ratings. For each model, 5,000 bootstrap iterations were performed. In each iteration, adjectives (and their phrase pairs) from the test set were resampled with replacement. For each selected adjective, the 15 individual human ratings for every phrase pair were also resampled with replacement before computing the mean rating per pair. Spearman's (ρ) was then calculated between these means and the corresponding model-predicted

cosine similarities. This was repeated independently for all adjectives in that iteration, and the resulting per-adjective correlations were averaged (unweighted) to yield a model-level score. The histograms in Figure 5.1 use 30 bins with fixed edges across plots for comparability. The shaded regions in each plot indicate the 95% percentile-based confidence interval, the dashed red line marks the bootstrap mean, and the dashed black line shows the raw correlation from the original ratings.

The close match between the bootstrap means and raw values indicates that resampling does not introduce bias, but instead captures the plausible range of results given variability in the human data. Consistent with the main results table, AT-Joint TSG achieves higher correlations than the unimodal baseline in both tasks, with its distributions shifted toward higher values supporting the major claims. The narrower spread for the multimodal model reflects more stable performance, while the broader spread for the baseline suggests greater sensitivity to sampling variation. For the remaining TSG models, we computed adjective-level bootstrap distributions. The results are provided in the Appendix C for reference.

5.3.5 Textual-Auditory Relationships

Some interesting observations can be made regarding textual-audio relationships as shown in the in Figure 5.2. The plot shows scores across various models for phrase learning.

It can be observed that both audio and textual similarity tasks exhibit a similar trend, with models performing well on both simultaneously. Text and audio modalities reinforce each other during learning. A model trained on both modalities can learn that adjectives like *loud* and *crunchy* modify nouns by embedding auditory properties into their representations. This alignment improves both audio and semantic predictions, leading to a more comprehensive understanding of sound-relevant language. Moreover, the annotations for audio similarity were collected based on how a human *perceives* the sound of a phrase rather than how they relate in meaning. The fact that models trained using multimodal embeddings can reproduce these human-like judgments demonstrates that the audio grounding enhances machine understanding of how humans perceive and categorise sounds. For example, while the actual sounds of a *creaky door* and a *creaky bridge* may differ due to environmental factors (e.g., echo or material differences),

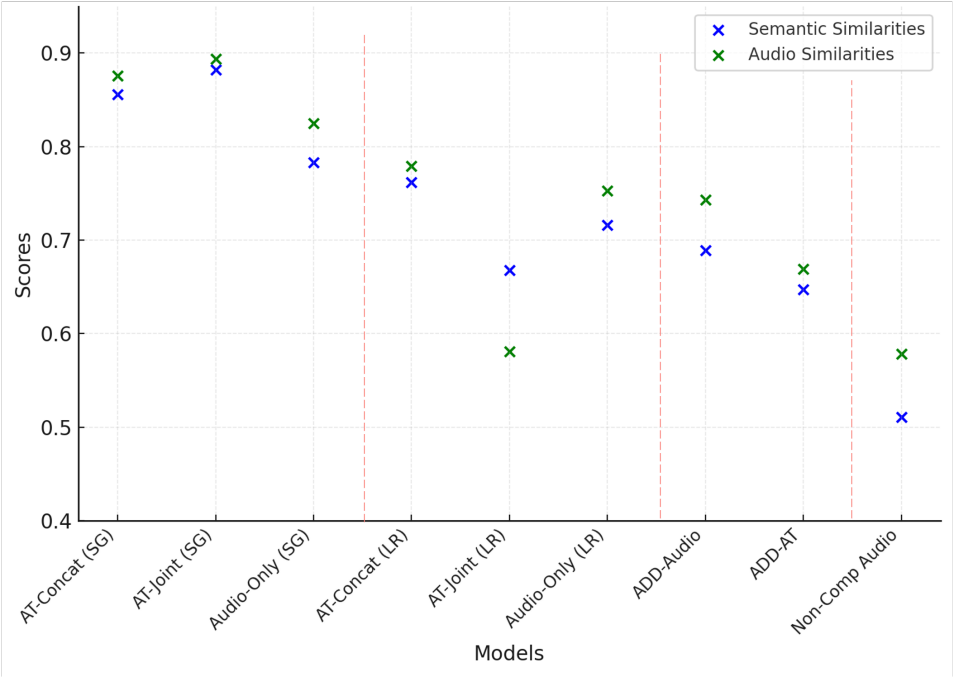


Figure 5.2: Scores across models with semantic similarities and audio similarities.

humans perceive both as *creaky* because of their shared auditory characteristics, leading to high perceived similarity in the annotations.

5.3.6 Multimodal Compositional Knowledge

The following sections further analyze the adjective-noun embeddings learned by both compositional and non-compositional models, assessing their ability to capture the compositional knowledge of phrases and exploring the impact of combining auditory information with textual data. The best-performing multimodal compositional model (AT-Joint) is compared with non-compositional models that rely solely on text or audio. This analysis follows a systematic approach, detailed below.

Data Preparation: For each model, all learned embeddings are first extracted. Embeddings from the training, validation, and test sets are then combined to form a unified dataset. This comprehensive integration is crucial for several reasons. *Firstly*, combining embeddings from all data splits prevents the evaluation from being biased by the limited size of the evaluation set alone. *Secondly*, this integration offers a clearer understanding of the model's overall performance and generalization across different contexts and examples.

5.3.6.1 K-Means Clustering

The next step involves creating clusters of adjective-noun phrases using k-means clustering. k-means is an unsupervised learning algorithm that partitions a set of data points into k clusters, where k is a predefined number. The algorithm works by minimizing the variance within each cluster, ensuring that data points within the same cluster are as similar as possible, while those in different clusters are distinct. Each cluster is represented by its centroid, which serves as the centre of the cluster and provides a reference point for grouping similar data.

Determining the Optimal k: The number of clusters (k) is a hyperparameter that needs to be determined. To select the optimal value of k, Silhouette method is used [112]. The Silhouette score measures the cohesion and separation of clusters, helping to identify the number of clusters that best represents the inherent structure of the data. This involves plotting silhouette scores for different values of k to find the point where the score is maximised. The score quantifies how similar a point is to its own cluster compared to other clusters. The value of k that yields the highest silhouette score is considered optimal.

Once the optimal k is determined, the phrase embeddings are clustered into k groups. Each cluster is then analyzed to identify common themes and characteristics, offering insights into the model's performance and the relationships it captures. The clusters are examined for both semantic and auditory groupings, providing an understanding of how the model integrates and differentiates these two types of information. For non-compositional text and audio models, we set $k = 4$, while for AT-Joint, we set $k = 3$.

Feature Scaling: Before clustering, phrase embeddings are normalised using a standard scaler . This process standardises the embeddings by removing the mean and scaling them to unit variance, bringing all features to the same scale. This prevents features with larger numerical ranges from dominating the analysis.

Similarity Computation: In this step, cosine similarity is used to evaluate the similarity between the query vector and all vectors within the same cluster. The similarity scores are calculated between the query phrase and each vector, then sorted to identify the top 10 most similar phrases. By prioritizing these, we concentrate on phrases that exhibit the

strongest semantic or auditory connections with the query phrase within the cluster.

5.3.6.2 Examples

Some examples from the analysis are presented in Figure 5.3. Grey rows indicate non-compositional audio and text-based similarities, while orange and blue highlight similar phrases for compositional audio and semantic similarities, using AT-Joint.

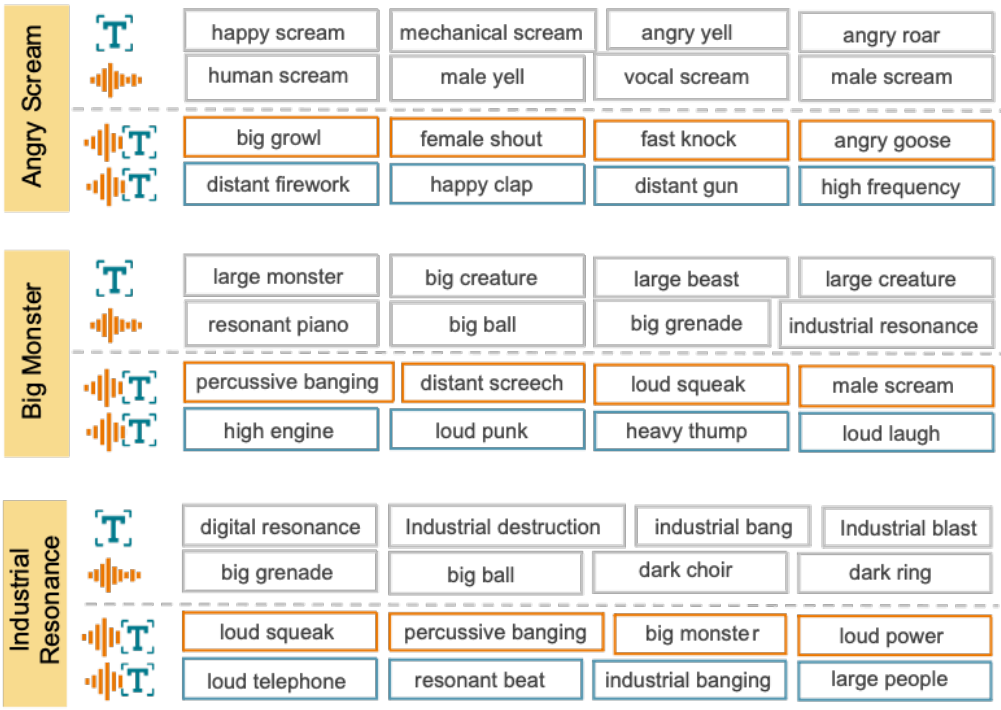


Figure 5.3: Query and its top 4 closely related phrases. Grey rows indicate non-comp audio and text-based similarities, while orange and blue signify similar phrases for compositional audio and semantic similarities, using AT-Joint.

Some examples from the analysis are presented in Figure 5.3. Grey rows indicate non-compositional audio and text-based similarities, while orange and blue highlight similar phrases for compositional audio and semantic similarities, using AT-Joint.

The examples suggest that text-only and audio-only models often produce predictions aligned more strongly with either semantic or auditory relevance, without always reflecting both aspects together. This can result in outputs that focus on literal sound matches or that diverge semantically from the intended concept. In contrast, the multimodal composition model appears to generate predictions that reflect elements of both meaning and sound.

For *Angry Scream*, the multimodal composition model outputs phrases such as *distant firework*, *distant gun*, and *high frequency*, which share some semantic and acoustic associations with the query. By comparison, the text-only model returns terms like *happy scream*, which may be semantically related but differ in sentiment, indicating a different emphasis in the type of similarity captured.

For *Big Monster*, the audio-only model produces results such as *resonant piano* and *big ball*, which, while sharing certain acoustic or lexical features, are less clearly connected to the sense of a large, imposing creature. The multimodal composition model includes outputs like *loud squeak* and *heavy thump*, which introduce elements potentially linked to both scale and sound.

For *Industrial Resonance*, the multimodal composition model suggests terms such as *percussive banging* and *loud telephone*, which relate to mechanical and resonant qualities. The audio-only model includes *Big Monster* among its predictions, while the text-only model suggests *industrial blast*. These examples show how different models can prioritise different aspects of similarity.

Overall, these observations point to differences in how models represent the relationship between semantic meaning and auditory information, with multimodal composition tending to include elements from both.

5.4 Conclusion

This chapter examines how different compositional models represent the relationship between text and audio. Across the semantic and audio similarity tasks, multimodal models using tensor skip-gram (TSG) often record higher Spearman correlation scores than audio-only and additive models. Approaches such as AT-Joint and AT-Concat, which combine text and audio, tend to produce embeddings that align more closely with phrase-level similarities. K-means clustering of adjective–noun embeddings shows patterns in which multimodal models group phrases in ways that reflect both meaning and sound, whereas single-modality models may emphasise one aspect more strongly than the other. These findings suggest that incorporating both text and audio with advanced composition methods can provide benefits for modelling phrase relationships.

Chapter 6

Multimodal Sentiment Analysis

*This chapter explores the application of **MultiCoDi** in sentiment analysis by integrating textual and auditory data in SST-5 dataset. It aims to capture emotional cues, such as tonal and auditory signals, that are often overlooked by text-only models.*

Sentiment analysis has long been a cornerstone of understanding opinions and emotions expressed in text, yet traditional approaches often fall short in capturing the full richness of human communication. Two critical gaps underlie this limitation: *first*, sentiment analysis techniques are often non-compositional, meaning that the phrase *not bad* would always be considered as a negative sentiment unless the two words are combined. *Second*, these approaches rely heavily on textual data while ignoring the multimodal nature of human expression, which incorporates diverse modes such as speech, audio, and visual cues to convey emotions more comprehensively. This chapter addresses these shortcomings by introducing MultiCoDi into sentiment analysis, combining textual data with auditory to capture subtle emotional cues often overlooked in purely textual approaches.

The chapter is organised as follows: First it begins with a review of relevant literature on sentiment analysis, covering levels of granularity, datasets, techniques, and the role of language compositionality. Next, it outlines the experimental setup, including data selection, preprocessing, and preparation, followed by the proposed methodology for integrating multimodal compositional embeddings into sentiment analysis. The chapter then presents the results and analysis, comparing the performance of multimodal and unimodal models and discussing key insights. It concludes with a summary of the

findings and their implications for enhancing sentiment analysis through multimodal approaches.

6.1 Literature

This section reviews sentiment granularity levels, datasets, traditional and compositional methods in sentiment analysis, highlighting recent advancements in the field.

6.1.1 Levels of Sentiment Analysis

Sentiment analysis aims to detect and understand the opinions or emotions expressed by individuals about particular topics, people, or entities. It can broadly be divided into four levels of granularity: document-level, sentence-level, phrase-level and aspect-level [113].

Document-level sentiment analysis evaluates the overall sentiment of an entire document, treating it as a unified entity that typically focuses on a single topic. The sentiment is categorised as positive or negative, making it particularly useful for understanding broad opinions expressed in reviews, articles, or reports. Studies, such as those by Pang et al. [114], Das and Chen [115], and Nongmeikapam et al. [116], have laid the groundwork for this approach by exploring methodologies to classify sentiment at the document level.

Sentence-level sentiment analysis identifies the sentiment within individual sentences, enabling more precise insights for applications like social media monitoring and review analysis. Early methods, such as [117], introduced a lexical-based approach to summarise product reviews, laid a foundation. Later, Socher et al. [118] advanced the field with recursive neural networks applied to the Stanford Sentiment Treebank, achieving fine-grained sentiment classification by capturing compositionality in sentence structures.

Phrase-level sentiment analysis focuses on identifying sentiment within specific phrases, capturing finer-grained emotional nuances that sentence-level analysis might overlook. For example, in the sentence *This laptop has a sleek design but poor performance*, phrase-level analysis can classify *sleek design* as positive and *poor performance* as negative. This approach is particularly valuable for handling mixed sentiments within a single sentence or document. Notable examples include the works of [119] and [46].

Aspect-level sentiment analysis identifies sentiments toward specific aspects on individ-

ual components within a text. For example, in the review, *The battery life of this phone is amazing, but the camera quality is disappointing*, aspect-level sentiment analysis detects positive sentiment for *battery life* and negative sentiment for *camera quality*. Significant studies in this field include works by [120] and [121].

6.1.2 Sentiment Analysis Techniques

Traditional sentiment analysis has evolved significantly over time, beginning with **lexicon-based methods**, that rely on predefined lists of words with associated sentiment scores to infer overall sentiment by aggregating these scores [122–124]. While these methods are interpretable and straightforward, they often fail to account for context, negations (e.g., *not good*), sarcasm, and evolving language patterns, limiting their applicability in complex real-world scenarios.

Machine learning methods such as Support Vector Machines (SVM) [125, 126] and Naive Bayes [127, 128], introduced a shift toward data-driven sentiment analysis by leveraging labeled data to detect sentiment patterns. These methods, while an improvement over lexicon-based approaches, rely heavily on manual feature engineering (e.g., n-grams, POS tags, syntactic dependencies), making them resource-intensive and less adaptable to new domains.

The introduction of **deep learning methods** marked a major breakthrough, allowing models to learn hierarchical representations of text directly from data without the need for manual feature extraction. Convolutional Neural Networks (CNNs) have been effective at capturing local features, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) excel at modelling sequential dependencies, making them suitable for complex sentiment analysis [129, 130]. Furthermore, transformer-based models like BERT have redefined the state of the art by capturing bidirectional context, enabling in-depth understanding of sentiment, including sarcasm, mixed sentiments, and complex structures ([8]).

Building on the success of these models, **multimodal sentiment analysis** emerged to address scenarios where sentiment is conveyed through multiple modalities, such as text, images, and audio. These approaches integrate diverse cues to create richer and more comprehensive sentiment representations, particularly in multimedia contexts like social media, reviews, and video content ([131, 132]). For example, multimodal models

can analyze how textual sentiment aligns with visual expressions or auditory tones to better understand user emotions. Chen et al. [48] pioneered by developing an image sentiment classifier that leverages object-based semantic concepts, achieving improved performance in multimedia contexts[1].

6.1.3 Compositional Sentiment Analysis

Traditional models that focus on individual words often miss the subtle ways in which sentiment is shaped by the composition of phrases and sentences. In their work, Moilanen et al. [45] introduced a model for sentiment analysis that mimics how meaning is derived in natural language by considering the structure and combination of words. Instead of merely counting positive and negative words, their approach accounts for how these words interact within a sentence to influence the overall sentiment. For instance, in the phrase *not happy*, the word *happy* alone would typically indicate a positive sentiment, but the addition of *not* reverses this sentiment, making it negative. Their model systematically combines these sentiments based on grammatical rules, leading to a more accurate sentiment analysis.

In another work, Yessenalina et al. [46] in 2011 proposed a research for phrase-level sentiment analysis that uses a compositional matrix-space model to capture complex semantic relationships. Building on earlier work by Baroni and Zamparelli [21], who represented adjectives as matrices and nouns as vectors, their method modeled all words as matrices and combines them through matrix multiplication. This approach, inspired by Rudolph and Giesbrecht [135], leverages the theoretical advantages of matrix-space models in accounting for word order and semantic nuances incorporating Ordered Logistic Regression to predict ordinal sentiment scores and introduces a unique training algorithm for this matrix-space model. Experimental results demonstrate significant performance improvements over traditional bag-of-words models on a standard sentiment corpus.

Later on in 2016, Kiritchenko and Mohammad [136] shifted focus to sentiment composition in phrases by creating a dataset of unigrams, bigrams, and trigrams containing both positive and negative words. They evaluated various learning algorithms and word

[1]For an in-depth exploration of advancements in sentiment analysis, refer to the comprehensive reviews by [133] and [134].

embeddings on this dataset to assess their performance across different linguistic patterns. In their study, they compiled a sentiment composition lexicon for phrases that include negators, modals, and adverbs, examining how these modifiers influence the overall sentiment of phrases [137] .

In 2017, another research based on compositional matrix-space model was proposed by Asaadi et al. [47] by introducing a two-step learning process aimed at improving the quality and computational efficiency of matrix-space models. Initially, the matrices are informed by unigram scores, which serves as the basis for a subsequent learning step that optimises relevant matrix entries using bigrams. This gradual learning method addresses the non-convex optimization challenges, ensuring better initialization and enhanced performance in sentiment composition. The model is then tested on fine-grained sentiment analysis tasks, demonstrating statistically significant improvements over traditional methods.

Adjective-noun pairs have been particularly influential in understanding sentiment. Chen et al. [48] developed an image sentiment classifier using adjective-noun pairs derived from image tags to detect objects and their attributes. Similarly, in 2021, Li et al. [49] created a visual sentiment prediction framework that translates images into textual descriptions, incorporating adjective-noun pairs for sentiment analysis. This framework uses a deep residual network and LSTM to generate initial descriptions, processes the text to retain key vocabulary, and embeds word vectors for training a sentiment prediction model. Borth et al. [138] introduced SentiBank, a large-scale visual sentiment ontology, using adjective-noun pairs like beautiful flowers or sad eyes extracted from YouTube videos and Flickr images to serve as mid-level descriptors for sentiment analysis.

Although the literature provides some evidences that compositional models and adjective-noun phrases are effective for sentiment analysis, their combined use remains limited. Additionally, the integration of this combination with multimodal information, particularly audio data, is rarely explored. This gap highlights the need for further research into the combined application of compositional models and adjective-noun phrases, especially within multimodal frameworks that incorporate audio cues for improved sentiment analysis.

6.1.4 Datasets

Several commonly used benchmarks in sentiment analysis are designed to evaluate and compare the performance of various models. Table 6.1 highlights some of the most widely recognised datasets:

Table 6.1: Common benchmarks for sentiment evaluation

Dataset	Reference	Primary Level	Domain	Polarity Scores
Amazon	[139]	Sentence & Aspect	Product Review	Binary
IMDB	[140]	Document	Movie Review	Binary
SST	[118]	Sentence & Phrase	Movie Review	Binary & Fine-grained
SemEval 2007	[141]	Aspect	Various	Binary & Fine-grained
CMU-MOSI	[142]	Utterance	Multimodal	Continuous

Amazon Product Review [139] contains reviews from various product categories on Amazon, offering a rich source of consumer opinions. It includes millions of reviews with detailed metadata such as product category, review text, and star ratings. The sentiment labels are typically binary (positive or negative), though some versions may provide finer-grained ratings from 1 to 5 stars. For binary classification, the dataset includes 3,600,000 samples for training and 400,000 samples for testing. In the 5-class classification version of the dataset, there are 3,000,000 training samples and 650,000 testing samples.

SemEval 2007 [141] consists of news headlines sourced from major outlets such as BBC, and Google News. This dataset is annotated with sentiment and emotion labels, capturing a broad range of emotional tones. It includes six primary emotions: anger, fear, disgust, sadness, surprise, and joy. Each headline is annotated not only with binary sentiment labels but also with fine-grained emotion ratings on a scale from 0 to 100, providing a nuanced view of emotional intensity.

IMDB 50k Movie Review [140] is a widely used collection of movie reviews sourced from the Internet Movie Database. It contains 50,000 reviews, evenly split between positive and negative sentiments. Each review is labeled with a binary sentiment score, reflecting overall positive or negative sentiment.

CMU-MOSI (Multimodal Opinion Sentiment and Emotion Intensity) [142] is a

large-scale resource for sentiment and emotion analysis, featuring over 23,500 video utterances from 1,000 YouTube speakers. The dataset includes balanced gender representation and covers diverse monologue topics. Each video is accurately transcribed and provides multimodal data (video, audio, and text). Sentiment is labeled on a continuous scale from -3 (highly negative) to +3 (highly positive).

SST (Stanford Sentiment Treebank) [118], derived from movie reviews, is designed for fine-grained sentiment analysis. The dataset parses sentences into individual phrases, each labeled with sentiment, enabling detailed analysis of sentiment at both the phrase and sentence levels. This hierarchical approach captures nuanced emotional variations that binary labels might miss. The SST dataset includes different versions: SST-2, which provides binary sentiment labels (positive and negative) for sentences, and SST-5, which offers a more detailed classification with five sentiment labels for a finer analysis. SST-2 includes around 12,000 movie reviews from Rotten Tomatoes, split into three distinct subsets: approximately 8,544 reviews for training, 1,101 reviews for development, and 1,658 reviews for testing. This dataset is annotated with binary sentiment labels, categorizing reviews as either positive or negative. The SST-5 dataset, an extension of SST-2, contains 11,855 sentences, with 215,154 unique phrases annotated for sentiment. SST-5 provides a more nuanced classification with five sentiment labels: very positive, positive, neutral, negative, and very negative. The dataset was created by parsing movie reviews into tree structures, where each node (phrase) in the tree is labeled with one of the five sentiment categories. This fine-grained approach enables detailed sentiment analysis, capturing a wider range of sentiment intensities within reviews.

6.2 Experimentation

This section outlines the experimental setup and procedures followed to evaluate the performance of the proposed multimodal sentiment analysis model. The experimentation is divided into data selection, data preprocessing and methodology.

6.2.1 Data Selection

The Stanford Sentiment Treebank (SST-5) dataset was selected for two key reasons. First, it provides fine-grained sentiment labels across five categories: very negative, negative,

neutral, positive, and very positive. This detailed classification facilitates a nuanced analysis of sentiments, capturing subtle emotional variations that binary labels might fail to discern. Second, SST-5 includes hierarchical annotations, offering sentiment labels not only for entire sentences but also for individual phrases within them. This supports precise sentiment analysis at multiple levels of granularity.

6.2.2 Data Preprocessing

SpaCy's POS Tagger was utilised to extract all adjective-noun phrases from the SST-5 dataset, which consists of 215,154 phrase entries. This process identified 82,664 phrases containing adjective-noun combinations, of which 504 phrases overlapped with audio data in the core dataset. The data was divided into training, testing, and validation splits, with 70% allocated for training and the remaining 30% evenly distributed between testing and validation. This resulted in 353 phrases for training, 76 for validation, and 75 for testing. A few examples are shown in Table 6.2.

Table 6.2: Filtered sentiments with selected phrases from SST-5 dataset.

Text	Label	Phrase
ranges from laugh-out-loud hilarious to wonder-what - time-it-is tedious.	neutral	loud laugh
between bursts of automatic gunfire, the story offers a trenchant critique of capitalism.	positive	automatic gunfire
rarely have i seen a film so willing to champion the fallibility of the human heart.	very positive	human heart
lucy 's a sad girl, that 's all.	negative	sad girl

Phrases from 27 multimodal adjectives were utilised, with 20 previously learned as part of the core dataset and 7 newly learned adjectives, including *young, automatic, busy, calm, crazy, long,* and *soft*. **Human annotations** for audio and semantic phrase similarities involving these adjectives were conducted by the authors. A total of 343 phrase pairs for the new adjectives were annotated, adhering to the guidelines outlined in Chapter 4.

6.2.3 Data Preparation

For this model, MultiCoDi phrase embeddings, developed for both semantic and audio similarity tasks as detailed in Chapter 3, were utilised. The data preparation process incorporated both audio and text data, ensuring compatibility and suitability for input into the neural network. For the **multimodal/audio data**, pretrained compositional audio features are standardised using z-score normalisation. This ensures that the features

have a mean of zero and a standard deviation of one, which helps in speeding up the convergence of the neural network. For the **text data**, the BERT tokeniser is used to tokenise the text. Each text string is prefixed with a special classification token [CLS], and a separator token [SEP] is added at the end. To maintain uniformity, text sequences are padded to a fixed length.

6.2.4 Implementation

A hybrid neural network model is integrated, leveraging both textual and auditory information for sentiment analysis. The model's architecture is designed to efficiently process and integrate textual and audio data for sentiment analysis. An overview of this architecture is shown in Figure 6.1.

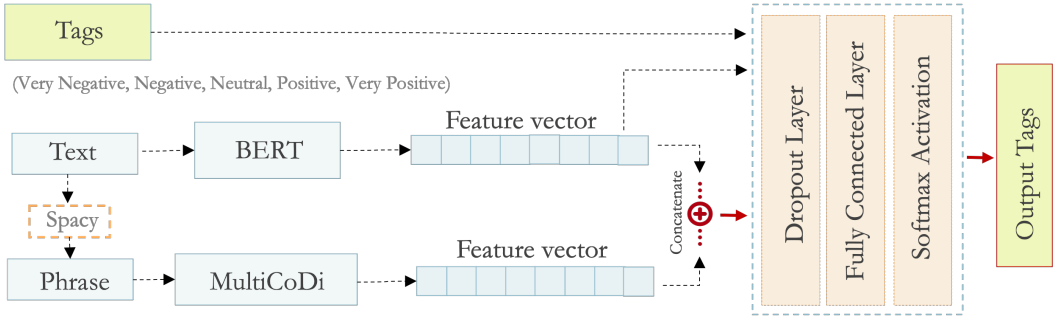


Figure 6.1: Multimodal sentiment analysis with compositional phrase embeddings.

Text Processing Component: The textual component centres around the BERT model, specifically the bert-large-uncased variant, which is known for its effectiveness in natural language processing tasks. It processes input text through multiple layers of transformer encoders that provide rich contextual embeddings. For this study, the outputs from the BERT model's pooling layer, which aggregates context-rich token embeddings into a single fixed-size embedding, are utilised.

Audio/Multimodal Processing Component: Parallel to the text processing, a multimodal/audio processing component is implemented using a simple neural network architecture. This component consists of a linear transformation layer, which maps the high-dimensional audio embeddings down to a lower-dimensional space of 256 units. This dimension reduction is crucial for aligning the audio data dimensionality with that of the textual data, facilitating their effective fusion. For instances where audio data is

not available or is deemed irrelevant, this component is not used, ensuring the model remains versatile and effective across varying data availability.

Feature Fusion and Classification: The fusion of text and multimodal/audio features is central to the model's design. The outputs from the BERT model and the multimodal/audio processing layer are concatenated to form a combined feature vector. This concatenated vector then passes through a dropout layer (nn.Dropout with a rate of 0.3), which helps prevent overfitting. Following the dropout layer, the combined features are fed into a final classifier, a linear layer that maps the combined features to the number of sentiment classes (five in this case).

Training Procedure: The training process for the hybrid neural network model was designed to optimise performance through a series of systematic steps. Data was loaded and batched using PyTorch's DataLoader, configured with a batch size of 32 to balance memory usage and computational efficiency, and included data shuffling to prevent the learning of unintended patterns. The training was conducted over 50 epochs, with an early stopping mechanism to halt training if validation performance ceased to improve, preventing overfitting. Loss was calculated using the Cross-Entropy Loss function, and model weights were updated through backpropagation using an Adam optimiser with an initial learning rate of 1e-5. A learning rate scheduler, ReduceLROnPlateau, was employed to adjust the learning rate based on validation loss. Performance metrics (loss and accuracy) were closely monitored and logged after each epoch to track progress and adjust parameters as needed.

6.3 Evaluation

6.3.1 Metrics

The models are evaluated based on the following evaluation metrics:

Accuracy: Accuracy measures how often the model makes correct predictions and is defined as the ratio of correctly predicted instances to the total number of instances, expressed by the following equation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

Where:

TP = True Positives $\quad$ TN = True Negatives

FP = False Positives $\quad$ FN = False Negatives

For this experiment, accuracy was chosen as the primary metric to evaluate the performance of the classification models. It was computed at the end of each epoch for training, validation, and test sets to monitor the model's performance over time.

Loss: Loss is monitored during training to quantify the difference between the predicted and actual values, where lower loss values indicate better model performance. For this classification task, the cross-entropy loss function is used, defined as:

$$\text{Cross-Entropy Loss} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{6.2}$$

Where:

N = Number of classes

y_i = Actual label (1 for the correct class, 0 for others)

$\hat{y}_i$ = Predicted probability of class i

The cross-entropy loss is calculated at each epoch by comparing the model's predicted probabilities with the actual class labels. It is computed for the training, validation, and test sets, offering a continuous measure of how well the model's predictions align with the ground truth. The loss increases as the predicted probability diverges from the actual label, thus providing a clear picture of model performance at each step of learning.

6.3.2 Results

Tables 6.3 and 6.4 report classification accuracies for the audio-relevant SST-5 phrase dataset, comparing multimodal and unimodal models for sentiment analysis. The majority-class baseline, added here for context, achieves 53.33%, reflecting the dataset's skew towards the "Neutral" class (53.33%), followed by "Negative" (29.33%), "Positive" (16.00%), and "Very Negative" (1.33%).

In the audio-derived setting (Table 6.3), AT-Concat with Tensor Skip-Gram (65.33%) achieves the highest accuracy, outperforming AT-Joint, Audio-Only, and non-compositional baselines. The difference relative to Non-Comp Text (62.70%) is

Table 6.3: Classification Accuracies for audio-relevant SST-5 phrase dataset using embeddings learnt via audio similarities

Model	Linear Regression	Tensor Skip-Gram
AT-Concat	62.67	65.33
AT-Joint	57.33	61.33
Audio-Only	54.67	58.67
ADD-Audio	55.33	
ADD-AT	54.50	
Non-Comp Audio	56.00	
Non-Comp Text	62.70	
Majority-Class Baseline	53.33	

Table 6.4: Classification Accuracies for audio-relevant SST-5 phrase dataset using embeddings learnt via semantic similarities

Model	Linear Regression	Tensor Skip-Gram
AT-Concat	53.33	64.00
AT-Joint	45.33	60.00
Audio-Only	56.00	58.67
ADD-Audio	54.00	
ADD-AT	52.70	
Non-Comp Audio	56.00	
Non-Comp Text	62.70	
Majority-Class Baseline	53.33	

modest but consistent. Both models use the same pre-trained BERT-based embeddings; the improvement for AT-Concat arises from combining these embeddings with compositional multimodal knowledge.

In the semantic-derived setting (Table 6.4), AT-Concat again achieves the highest TSG score (64.00 %), followed by AT-Joint (60.00 %) and Audio-Only (58.67 %). As with the audio-derived setting, Non-Comp Text (62.70 %) remains competitive, showing that purely textual models already carry strong sentiment cues.

Across both tasks, Tensor Skip-Gram generally outperforms Linear Regression, suggesting that TSG's ability to model more complex interactions between adjective–noun components benefits classification. Still, the advantage of multimodal over unimodal approaches is not uniform: the clearest and most consistent gains are for AT-Concat

(TSG) over unimodal baselines, but these gains are small relative to Non-Comp Text.

6.3.3 Analysis

To assess robustness for key comparisons, non-parametric bootstrap resampling was applied to the test set predictions of four selected models: AT-Concat (TSG) and their closest unimodal baselines (Non-Comp Text for semantic embeddings, Non-Comp Audio for audio embeddings). For each model, accuracies were recalculated over 5,000 resamples, and 95 % confidence intervals (CIs) were estimated from the resulting distributions (Figure 6.2).

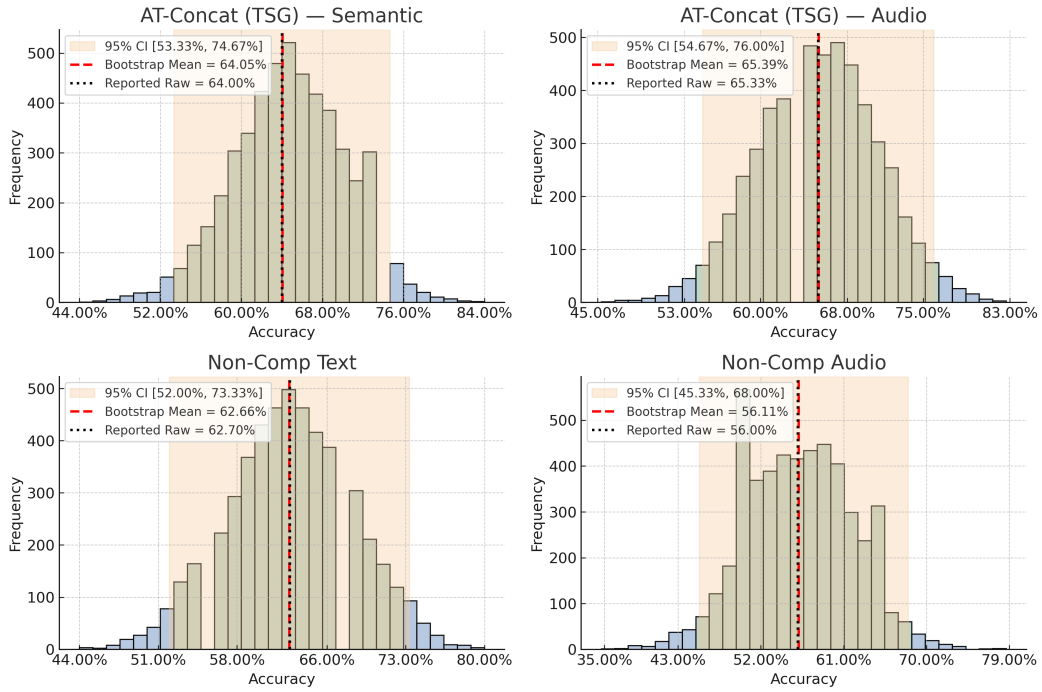


Figure 6.2: Bootstrap accuracy distributions (5,000 resamples) for selected multimodal and unimodal models on the audio-relevant SST-5 test set. Top row: AT-Concat (TSG) — Semantic vs. Non-Comp Text. Bottom row: AT-Concat (TSG) — Audio vs. Non-Comp Audio. Shaded areas indicate 95% confidence intervals (CIs); dashed red lines mark bootstrap means; dotted black lines show reported raw accuracies.

In the audio-derived setting, AT-Concat (TSG) shows a clear separation from the Non-Comp Audio baseline, with minimal CI overlap, indicating a statistically reliable advantage. In the semantic setting, the CI for AT-Concat (TSG) partially overlaps with that of Non-Comp Text, suggesting that the difference is not statistically significant at the 95 %

level. This limited separation is likely due to the small and class-imbalanced nature of the audio-relevant test set, which constrains statistical power.

Despite this, the distributions for AT-Concat are consistently shifted to the right relative to their unimodal counterparts across thousands of resamples. This pattern supports a genuine performance trend, which may become statistically significant with a larger and more balanced evaluation set.

6.3.4 Examples

Table 6.5 compares multimodal compositional and non-compositional models. Sentiment labels are denoted as Negative (-), Very Negative (–), Neutral (=), Positive (+), and Very Positive (++). The *True* column shows the gold sentiment labels, the *Text* column reflects human-expressed sentiments, and sound-relevant phrases are listed in the *Phrase* column. The *Perf.* column provides qualitative ratings: *Strong*, *Weak*, or *Neutral*.

Table 6.5: Performance comparison of NC(Non-Comp) Text, AT-Concat (ATC) models using semantic (Sem) and auditory (Aud) embeddings.

True	Text	Phrase	NC Text	ATC (Sem)	ATC (Aud)	Perf.
+	the movie has lots of dancing and melodious music	melodious music	+	+	++	Strong
=	84 minutes of rolling musical beat and supercharged cartoon warfare.	musical beat	-	=	=	Strong
-	the entire movie is about a boring, sad man being boring and sad.	sad man	-	-	=	Neutral
-	writer-director randall wallace has bitten off more than he or anyone else could chew, and his movie veers like a drunken driver through heavy traffic.	heavy traffic	- -	- -	=	Weak

The table shows a few examples comparing Non-Comp Text with *AT-Concat* using semantic *(Sem)* and auditory *(Aud)* embeddings. *AT-Concat (Aud)* tends to perform well when phrases have clear audio associations (e.g., melodious music), suggesting some benefit from including audio-based information. For phrases with less distinct sound cues (e.g., sad man), the advantage is smaller. More complex phrases (e.g., heavy traffic) remain challenging, likely due to the difficulty of modelling multiple interacting components. These examples illustrate how multimodal compositional models can offer advantages in certain contexts, while also highlighting their current limitations.

6.4 Conclusion

This chapter examined the role of compositional knowledge and multimodal information in sentiment analysis. In these experiments, models combining audio and textual data, particularly those using TSG, sometimes achieved higher accuracies than single-modality baselines, though gains were not consistent across all settings. These findings indicate that multimodal integration can offer benefits in certain cases, but the extent of improvement depends on the task and data characteristics. The next chapter examines the application of multimodal distributional semantics to recommendation systems.

Chapter 7

Multimodal Recommendations

This chapter explores the application of multimodal distributional semantics in recommender systems by integrating textual, auditory, and visual modalities, using a weekly BBC TV programs dataset as a case study.

Over the past few decades, classical recommender systems have focused on developing techniques to help users navigate the overwhelming volume of video content available online. These methods fall into three primary categories: collaborative filtering (CF) [143], content-based filtering (CBF) [144], and hybrid approaches [145]. Collaborative filtering relies on users' historical behavior to generate recommendations but often struggles with the cold start problem, where insufficient data on new items or users limits its effectiveness. In contrast, content-based filtering utilises the semantic similarity of item content, typically text. Hybrid approaches aim to address these limitations by combining the strengths of CF and CBF.

Building on these foundations, modern recommendation systems have increasingly adopted vector semantics, where the content of words and documents is represented as high-dimensional vectors [146, 147]. Advances in natural language processing have further refined these representations through neural network models such as Word2Vec and Doc2Vec [6]. These vector-based approaches have paved the way for richer content understanding by incorporating additional layers of information, including audiovisual and cognitive features, as seen in [11, 33].

Recently, multimodal recommender systems have emerged as a significant advancement over classical systems by integrating diverse data modalities such as text, audio, and

video alongside user ratings to deliver more personalised and accurate recommendations. For instance, Zhu et.al. [148] and Barkan et.al. [149] have explored this integration, though many existing systems remain limited. Some, like Yang et.al. [50], only consider tags and titles as textual data, while others, such as Ekenel et.al. [51], combine images with tags. Bougiatiotis & Giannakopoulos [52] took it a step further by integrating audio and video with subtitles but still falls short by ignoring genres, ultimately failing to outperform metadata-only systems.

This chapter aims to address these shortcomings by building on the multimodal approach proposed by Kiela & Clark [33] (discussed in Chapter 2) and extending it from word-level representations to documents, further enriched with genre and visual vectors. The aim is to enhance both the precision and diversity of recommendations, offering users suggestions that are not only highly relevant but also varied and engaging. Unlike the existing BBC recommender system, which relies heavily on genre-based recommendations and tends to lack diversity by repeatedly suggesting programmes from the same categories, this approach integrates auditory and visual relevance to deliver a richer and more varied viewing experience.

This chapter is structured as follows: Section 7.1 provides an overview of the BBC dataset used in the study. Section 7.2 details the methodology, discussing each modality and the fusion process. Section 7.3 presents the evaluation techniques and results. Finally, Section 7.4 present the final conclusions and outline future directions. For additional information, see Appendix D.

7.1 BBC TV Programmes Dataset

The dataset used for training the multimodal content recommendations consists of 145 unique BBC TV programmes, organised in a hierarchical structure to allow for a thorough evaluation. This hierarchy categorises the programmes into episodes, series, and brands. At the top of this structure is the Top-Level Editorial Object (TLEO), representing the highest level of classification, ensuring that each entry is distinct. Programmes in the dataset may belong to a series, a brand, or exist as standalone episodes. For example, *EastEnders* is classified as a brand without a series designation. Each TLEO corresponds to one unique episode, resulting in 145 distinct episodes in the dataset. Alongside the programme content, the dataset includes comprehensive metadata such as genre, format,

service, titles, descriptions, subtitles, audios, and videos, which provide a rich set of information for developing and evaluating the recommendation system[1].

7.2 Recommendation Framework

The proposed recommendation framework integrates textual, audio, and visual content to improve recommendation quality. As depicted in Figure 7.1, each modality is independently processed to construct its respective similarity matrix.

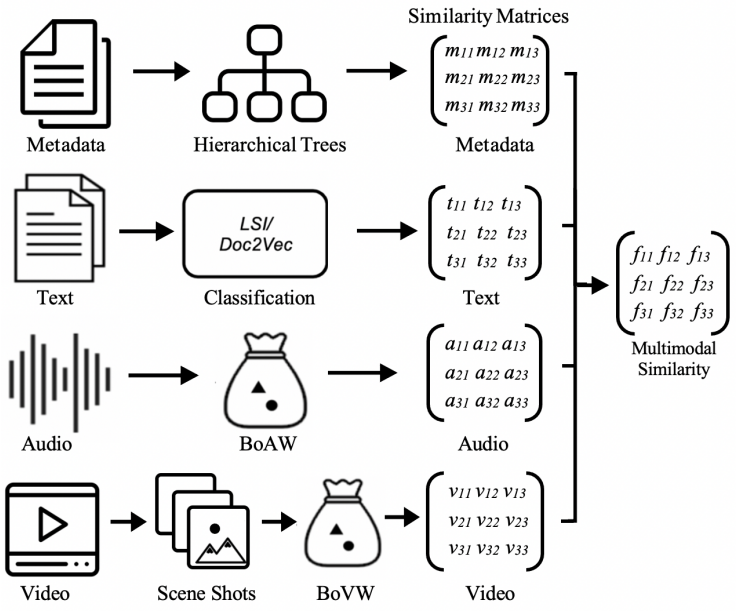


Figure 7.1: Methodology of the multimodal content recommendation framework.

Text is analyzed using Latent Semantic Indexing (LSI) or Doc2Vec embeddings, audio features are extracted via the Bag of Audio Words (BoAW) model, and video content is processed using the Bag of Visual Words (BoVW) model, with scene shots analyzed to create the video similarity matrix. The modality-specific similarity matrices are subsequently combined to form a comprehensive multimodal similarity matrix. By leveraging the strengths of each content type—text, audio, and video—this unified matrix delivers more holistic and accurate recommendations, capturing the diverse elements of the programmes. This integration significantly enhances the precision and relevance of the recommendation system.

[1]The BBC dataset was chosen as it includes all major modes of information—text, audio, video, and genres—alongside user viewing data, unlike publicly available datasets such as MM-IMDb [150], which lacks subtitles, and YouTube 8M [92], which is limited to visual features.

7.2.1 Textual Recommendations

Textual recommendations encompass the textual content associated with programmes, which typically includes subtitles and metadata entities such as genres and formats. Genres provide significant information about the category of a program; for instance, the genre *Documentary* represents all programs falling under the documentary type. Traditionally, genres are used in recommendations based on the assumption that if a person likes a certain type of program, like news, they will likely be interested in other similar programs. However, human preferences are not always consistent, and people often seek variety out of curiosity, desiring different but still relevant recommendations. For this experiment, the focus is on subtitles, as they contain substantial semantic information about a program's content and theme.

7.2.1.1 Subtitle Vectorization

Latent Semantic Indexing (LSI) [151], a topic modelling technique, is applied to extract data from subtitles. LSI is a two-step procedure. Firstly, a document- term matrix is generated via a low-rank approximation obtained from the term vector space projections of the Bag of Words vectors. Secondly, Singular Value Decomposition (SVD) is applied to the document-term matrix, where the newly created eigenvectors represent the concepts within the latent space. We worked with 50 dimensional spaces. LSI improves on the term-document matrices, but does not take word order into account. To deal with this, we worked with neural semantics embeddings Doc2vec [152]. Doc2vec is an extension of the neural semantic word embeddings Word2vec [6]. We worked with Paragraph Vector Distributed Memory (PV- DM), which concatenates the unique document ID with the context words with respect to the specified context window over the text and preserves the order of words.

7.2.1.2 Attributes (Genres)

These representations are based on editorially-assigned attributes of programmes. Each programme has a genre which is hierarchical with up to three levels (e.g. *factual*, *factual/sci&nature*, *factual/sci&nature/nature&env*) and a match can occur at any level. The hierarchical structure is broken down into a set of attributes by traversing the tree. This set is represented by vectors, where each column represents a genre subtree obtained

from a partial tree of the genre hierarchy and each column entry is a binary value denoting the relation between the program and the genre, i.e. whether the programme had that partial tree as part of its genre hierarchy. Using the vectors thus obtained, we computed a metadata similarity matrix, where a complete match receives a score of 1 but the score is halved for each level above.

7.2.2 Audio Recommendations

Audio Preprocessing: To prepare programme audios in the dataset for feature extraction, unwanted data is carefully removed, including silences, episode introductions, news segments (e.g., 90-second updates), advertisements, trailers, and previews for upcoming episodes. This step ensures that only the most pertinent audio content is retained for subsequent analysis, enhancing the quality and relevance of the extracted features. The overall method is shown in Figure 7.2.

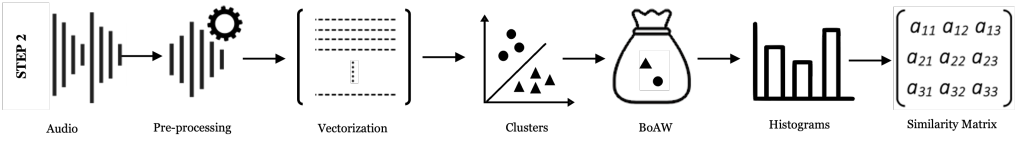


Figure 7.2: Proposed method for generating auditory recommendations

Vectorization: Since the audio data comprises a mix of voice and music, five types of low-level acoustic features are applied, which are widely recognised in both speech recognition and music information retrieval tasks.

(a) *Mel-Frequency Cepstral Coefficients (MFCC)* estimate the short-term power spectrum of a sound and are widely used in sound analysis due to their alignment with human auditory perception.

(b) *Spectral Centroid* is the balance point or the midpoint of the spectral energy distribution/spectrum of a sound. It gives an indication of the “brightness" of a sound which in turn is an approximation of high-frequency content in a signal.

(c) *Zero Crossing Rate (ZCR)* is the rate at which the signal crosses the zero point and changes itself from positive to negative or negative to positive.

(d) *Spectral Flatness (SF)* quantifies how noise-like or tone-like a signal is by assessing the number of peaks or resonant structure in a spectrum. Values near 1.0 indicate a

flat spectrum with similar amount of power in all spectral bands, as seen in white noise.

(e) *Root Mean Square (RMS)* calculates the root-mean-square of a signal. For a digitised signal, it can be calculated by squaring each value, finding mean and taking the square root of the result. In terms of audio signals it represents the average *power* of a signal.

To extract these features, the LibROSA library [153], a robust Python package for music and audio analysis, is utilised. The chosen acoustic features are extracted, concatenated, and combined into comprehensive audio feature vectors.

7.2.2.1 BoAW

The bag-of-Audio-Words (BoAW) approach is inspired by bag-of-words (BoW) in text mining. The difference is that in text mining there are textual words to create a word dictionary while in audio signal processing, these textual words are replaced by an audio word which corresponds to a combination of acoustic features. BoAW involves extracting low-level audio features such as MFCCs from audio signals. These features are clustered into a fixed number of groups, known as a *codebook*, using techniques like k-means clustering, where each cluster centre represents an *audio word*. The audio signal is then represented as a histogram of these audio words, counting the occurrences of each cluster across the signal. This approach converts variable-length audio signals into fixed-length vectors. BoAW has successfully applied to many audio information retrieval and recognition tasks like multimedia event detection [154–156] and acoustic event detection [157–159].

For this project, the dictionary derived from the BoAW is used to generate histograms representing audio word distributions for all audio files. BoAW is implemented using k-means classifier with $k = 50$, resulting in 50 audio words per file. To analyze relationships between audio files, a similarity matrix is constructed by computing cosine similarities between these histograms.

7.2.3 Visual Recommendations

The visual recommendations follow the same set of step outlined in Figure 7.2, tailored specifically for visual data.

Video Preprocessing: The video content for each programme is represented by a sequence of still images extracted from the middle of each scene. This extraction is crucial for capturing the most representative visual features while minimizing motion blur. The SceneDetect application from the PySceneDetect library was used for this task, employing the ContentDetector algorithm. The algorithm was configured with a detection threshold of 30 and a minimum scene length of 15 frames to accurately identify distinct scenes based on changes in visual content. By selecting images from the middle of each scene, the process ensures that these frames are the most indicative of the scene's content. For each programme, a carefully selected subset of 600 images was chosen to balance comprehensive scene representation with computational efficiency.

Vectorization: After extracting the still images, the next step involves feature vectorization to quantify the visual content of each scene. The Scale-Invariant Feature Transform (SIFT) descriptors were selected for this purpose. SIFT is a powerful tool in image processing, known for its robustness in image matching, as well as its effectiveness in object detection and recognition. Each keypoint in an image is characterised by a 128-dimensional feature vector generated by SIFT. These vectors capture essential details about the visual content, making them highly suitable for the subsequent classification tasks. The decision to use SIFT was guided by its proven performance in the literature [160–162].

7.2.3.1 BoVW

The Bag of Visual Words (BoVW) model quantises the SIFT descriptors by clustering them into visual words, thereby creating a vocabulary that represents the visual content of each programme. Similar to BoAW, this quantization process was executed using K-means clustering, with the number of clusters $k = 300$. Each cluster represents a *visual word*, and together, these words form a vocabulary that encapsulates the essential visual features of the programme. The resulting visual word vocabulary allows for efficient scene classification and contributes significantly to the overall accuracy and effectiveness of the multimodal recommendations.

7.2.4 Fusion

Following the multimodal fusion techniques discussed in Chapter 2, late fusion was employed in this project as the primary approach due to its compatibility with the nature of the data. Visual, audio, and textual modalities were processed independently, and their contributions were combined through weighted addition of similarity matrices derived from genres, LSiI, Doc2Vec, audio, and video data. This approach preserved the unique characteristics of each modality while enabling a seamless combination at the decision level, ultimately enhancing the accuracy and robustness of the recommendations[2].

7.3 Evaluation and Results

Evaluation Method: A personalised recommender evaluation system based on the MyMediaLite library [163] was utilised to assess the performance of the representations. This system processes binary user-item preference data for training and testing, obtained from BBC iPlayer media server logs. A user's positive preference is recorded when their viewing time exceeds 5 minutes, a threshold determined by observing the lapse rate (the rate at which users stop watching a programme). The first week of recorded data serves as the training set, while a subset from the following week is used for testing. The training data comprises 1,390,540 viewings from 33,958 users across 145 TV programmes, while the testing data includes 47,707 viewings from 10,000 users and 141 programmes, with the test users being a subset of those in the training data.

To generate recommendations, the Weighted Item-based K Nearest Neighbours (KNN) algorithm provided by MyMediaLite [163] was applied. In this setup, programme similarity is based on overlapping viewing histories: two programmes are considered similar if they have been watched by many of the same users. The resulting programme–programme similarity matrix, built directly from viewing patterns, provides what we refer to as a **user-based** model. This model reflects actual audience behaviour and serves as a reference for evaluating the content-based recommendations. In practice, the KNN approach ranks each programme's nearest neighbours according to their similarity scores, and the

[2]Early fusion was initially explored, combining feature vectors from audio, text, and genres using operations like addition, multiplication, and averaging. However, this approach proved less effective, primarily due to the sparse and uneven distribution of genre feature vectors, which negatively impacted the system's performance when integrated at the feature level.

top-N most similar items are then recommended to a user based on the programmes they have already watched. This allows the system to recommend items that are most closely related, according to historical audience behaviour, to those the user has previously engaged with.

Accuracy was measured using Mean Average Precision (MAP), which evaluates the number of correctly predicted viewings found in the top-N recommendations (hits). Additionally, Intra-list Diversity (ILD) was computed to measure the genre diversity within the recommendations for each individual user. The representations were evaluated using both individual and fused models.

Results: The results of the evaluations across different modalities and their combinations are detailed in Tables 7.1 through 7.5. Each table provides insights into the performance of textual (LSI, DM), audio (A), video (V), and genre (G) models, both individually and in various fused combinations. The weights associated with each modality in the fused models indicate the extent of their contribution to the overall performance.

Table 7.1 presents the performance metrics for **individual modalities**. The textual models, Doc2Vec (DM) and Latent Semantic Indexing (LSI), outperform the genre model (G) in both Mean Average Precision (MAP) and Intra-List Diversity (ILD). Specifically, Doc2Vec achieves a MAP@10 of 11.76% with an ILD@20 of 80.37%, while LSI scores 11.30% in MAP@10 and 76.69% in ILD@20. Despite having lower MAP scores, audio and video modalities contribute significantly to diversifying the recommendations, with the video modality achieving the highest ILD@20 of 82.05%, though its MAP@10 is the lowest at 3.88%. This indicates that while audio and video may not excel in precision, they are valuable for enhancing the diversity of recommendations.

Table 7.2 introduces the **text-only fusion** model, combining genre information with both Doc2Vec and LSI embeddings. This configuration achieves a MAP@20 of 15.20% and an ILD@20 of 71.90%, surpassing the performance of any individual modality. The results indicate that combining neural and one topic-based representations, along with structured genre metadata, contributes meaningfully to both accuracy and diversity. While this model does not rely on audio or visual inputs, it still provides strong performance, demonstrating the strength of semantic features extracted from subtitles alone.

Table 7.3 shows the performance of fused models that combine **textual, audio, and**

Table 7.1: Singular model evaluations

Model	MAP@10	ILD@10	MAP@20	ILD@20
Genre (G)	10.78	35.52	12.77	52.72
Doc2vec (DM)	11.76	77.20	13.88	80.37
LSI	11.30	69.89	13.40	76.69
Audios (A)	6.67	77.96	8.11	81.38
Videos (V)	3.88	81.43	4.97	82.05
User-Based	15.60	79.73	18.51	80.90

Table 7.2: Fused textual-only evaluations

Model	MAP@10	ILD@10	MAP@20	ILD@20
G + DM+ LSI 1.2 0.8 0.6	13.40	64.80	15.20	71.90

Table 7.3: Fused textual, audio and genre evaluations

Model	MAP@10	ILD@10	MAP@20	ILD@20
LSI+ A+ G 0.5 0.3 0.2	12.87	63.10	15.21	71.65
DM+ A+ G 0.7 0.2 0.1	13.78	59.03	16.17	67.87
LSI+ DM+ A+ G 0.7 1.5 0.2 0.65	14.98	61.29	17.45	70.00

Table 7.4: Fused textual, video and genre evaluations

Model	MAP@10	ILD@10	MAP@20	ILD@20
LSI+ V+ G 1.00 0.13 1.00	13.48	53.00	15.74	64.20
DM+ V+ G 1.8 0.1 1.00	14.23	54.75	16.62	64.68
LSI+ DM+ V+ G 0.7 1.5 0.12 0.65	14.99	60.62	17.45	69.58

Table 7.5: Fused textual, audio, video, and genre evaluations

Model	MAP@10	ILD@10	MAP@20	ILD@20
LSI+ DM+ A + V+ G 0.7 1.5 0.12 0.1 0.65	15.07	61.17	17.55	69.88
User-Based	15.60	79.73	18.51	80.90

genre modalities. The top-performing model in this group combines LSI, DM, audio, and genre (with respective weights of 0.7, 1.5, 0.2, 0.65), achieving a MAP@10 of 14.98% and an ILD@20 of 70.00%. This fusion improves MAP over individual models while preserving diversity, showing that combining modalities boosts both.

Table 7.4 highlights the performance of models that fuse **textual, video, and genre** modalities. The best result in this category comes from combining LSI, DM, video, and genre (with weights 0.7, 1.5, 0.12, 0.65), resulting in a MAP@10 of 14.99% and an ILD@20 of 69.58%. This combination provides a balanced approach, improving both MAP and ILD, although it does not quite reach the performance of the user-based model.

Table 7.5 presents a fully fused model integrating all modalities: **LSI, DM, audio, video, and genre**. This comprehensive fusion achieves the highest MAP@10 of 15.07% and an ILD@20 of 69.88%. These results are closely aligned with the user-based model, which has a MAP@10 of 15.60% and an ILD@20 of 80.90%, demonstrating the effectiveness of combining multiple modalities to closely estimate user preferences and behaviors.

Results demonstrate that splitting an episode's content into textual, auditory, and visual levels significantly enhances the quality of recommendations. Learning subtitle features separately benefits the system by capturing textual semantic information, where episodes using similar English words are more closely related. Auditory semantics further refine the recommendations by distinguishing whether an audio is more noise-like or tone-like, allowing the system to avoid recommending a programme with intense music and screeching voices to a user who prefers hushed tones. Adding the visual aspect provides another layer of understanding by analyzing the visual content of the episodes. This allows the system to compare episodes based on visual similarities, such as colour schemes, scene compositions, or even the presence of specific visual motifs. For instance, an episode with dark, suspenseful imagery might be less recommended to user who prefers bright, light-hearted visuals.

7.4 Conclusion

In this chapter, a multimodal content recommendation system for BBC TV programmes was developed. The developed system demonstrates a significant advancement over traditional genre-based methods, showcasing the potential of multimodal integration

in enhancing the relevance and diversity of TV programme recommendations. Results demonstrate that splitting an episode's content into textual, auditory, and visual levels significantly enhances the quality of recommendations. In the future, the weekly dataset could be expanded to include a wider range of programs collected over several weeks or months, enabling more comprehensive training and evaluation of the system. Additionally, future iterations could integrate MultiCoDi to enhance the interpretation of emotional tones across diverse content.

Chapter 8

Conclusion and Future

8.0.1 Summary

Language is a rich and dynamic medium that encompasses both context and perception. This thesis embraced that complexity by exploring multimodality in language compositions, bridging the gap between purely textual representations and real-world sensory grounding. Its core contribution is the development of MultiCoDi, a framework that grounds compositional models in auditory data. Inspired by tensor-based compositional frameworks and type-driven approaches, MultiCoDi integrates noun vectors and adjective matrices learned from both textual and auditory modalities. Unlike previous work that mainly focused on unimodal settings or visual grounding, this research demonstrated that grounding linguistic representations in sound provides a richer understanding of language.

Another key contribution is the creation of a sound-relevant phrase similarity dataset, bridging a gap in benchmarks by evaluating phrases across semantic and auditory dimensions. The dataset played a central role in evaluating compositional models on tasks where sound is critical, showing that multimodal models consistently outperform unimodal baselines. Matrix-based compositions, in particular, proved to be more effective than vector-based approaches, highlighting the importance of capturing the relationships between linguistic components.

This thesis also demonstrated the practical applications of MultiCoDi in compositional sentiment analysis. By combining textual and auditory features in a compositional manner, MultiCoDi effectively captured subtle shifts in sentiment that traditional models

often miss. Additionally, a separate multimodal distributional framework was developed as a baseline for future experiments in a content recommendation system for TV programmes. By integrating auditory, textual, and visual features, this system achieved more precise and diverse recommendations.

The findings highlight the benefits of integrating sensory modalities with textual representations to improve linguistic meaning. Grounding phrases in both text and sound provides a more comprehensive and accurate understanding of language. The success of matrix-based compositions further underscores the importance of capturing the relationships between different components of language, making the resulting representations contextually rich and precise. Despite these advancements, this thesis acknowledges several limitations. The lack of extensive multimodal datasets covering various linguistic structures and auditory contexts constrained further experimentations. Additionally, the performance of multimodal models may vary across domains due to differences in auditory-textual feature interactions in different contexts.

8.0.2 Future Directions

Looking ahead, there are several promising directions for future research.

Extension to Other Modalities: The proposed framework can further be expanded by incorporating visual data, such as images or videos (a snippet of which has been shown in recommendations). Prominent supporting works on images include works by Bruni et al. [11], Kiela & Clark [33] for words, and Lewis et al. [37] and Wazni et al. [38] for compositional models. The proposed MultiCoDi framework can be extended to images as shown in Figure 8.1. More precisely, for example for the proposed concatenated tensor

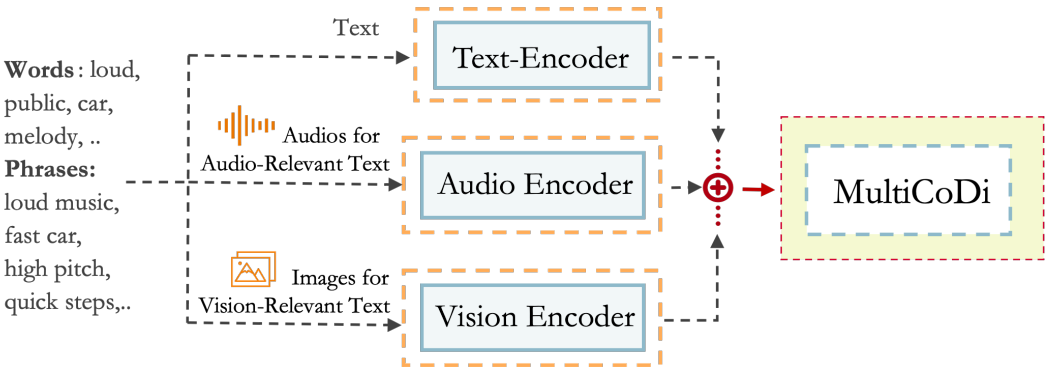


Figure 8.1: Example MultiCoDi extension to vision.

Skipgram, the objective function would become:

$$\sum_{(\boldsymbol{c'^a},\boldsymbol{c'^t},\boldsymbol{c'^v})\in\mathcal{C}^a\times\mathcal{C}^t\times\mathcal{C}^v} \log\sigma\left(\mathbf{A}\langle\boldsymbol{n^a},\boldsymbol{n^t},\boldsymbol{n^v}\rangle\cdot\langle\boldsymbol{c'^a},\boldsymbol{c'^t},\boldsymbol{c'^v}\rangle\right)$$

Here, $\langle\boldsymbol{n}^a,\boldsymbol{n}^t,\boldsymbol{n}^v\rangle$ represents the concatenation of the fixed pre-trained audio, textual, and visual embeddings of a noun, and $\mathcal{C}^a$, $\mathcal{C}^t$, and $\mathcal{C}^v$ are the sets of positive and negative contexts of the adjective-noun phrase. Positive contexts can be learned through multiple auditory or visual representations of the same entity or by leveraging large textual corpora (such as UKWaC). Moreover, the proposed dataset can also be utilised for this audio-visual analysis. Notably, several phrases within the dataset exhibit an auditory-visual overlap, providing an opportunity for cross-modal exploration. Examples of such phrases include *fast car, angry girl, fast food, angry monster, sad man, happy person, distant blast, big drone, and big door.*

Extension to Other Language Structures: Future work could extend the proposed framework to more complex syntactic structures, such as subject-verb-object (SVO) triplets or verb-phrase combinations. Several promising studies support this direction including Wijnholds & Sadrzadeh [24] have explored learning representations for transitive verbs and other functional types using multilinear maps, applying Combinatory Categorial Grammar (CCG) for type-driven composition and Wazni et al. [38] extended it by grounding verb matrices with visual data through linear regression. Future work could involve extending MultiCoDi to SVO triplets. Tensor-based operations, such as Copy-Subject or Copy-Object, could be used to compute the SVO embedding. For instance, the Copy-Subject operation can be formalised as:

$$\overrightarrow{\text{subj verb obj}} = \overrightarrow{\text{subj}} \odot \left(\overline{\text{verb}} \times \overrightarrow{\text{obj}}\right)$$

Here, the subject and object will be represented as grounded vectors, while the verb as a grounded matrix. For example, for the sentence, *The dog barks at the cat*, the subject (*dog*) and object (*cat*) could be represented using audio embeddings, while the verb (*barks*) would be learned to reflect the dynamic nature of the action through context-aware representations.

Audio Captioning and Multimodal Applications: The proposed framework can be

applied to various textual-audio understanding tasks, including automated audio captioning. The work of Eren et al. [164] supports this, demonstrating that incorporating partial captions alongside audio inputs can enhance the performance of audio captioning models.

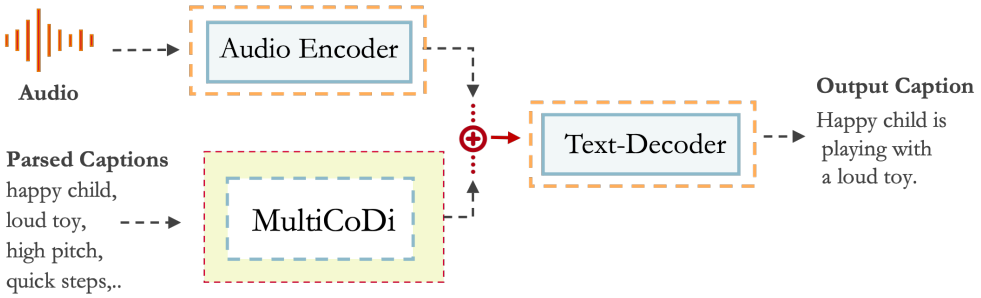


Figure 8.2: Example application of MultiCoDi in automated audio captioning.

Figure 8.2 illustrates a practical example of how the proposed framework can be integrated into audio captioning systems. In this task, where the objective is to generate descriptive text based on auditory inputs, the use of multimodal language compositions can enhance the contextual understanding of audio events, leading to more accurate and contextually relevant captions.

Appendix A

Modelling Multimodal Phrases

A.1 Textual Composition

We implemented a text-only approach to learning adjective-noun phrase representations, building on the compositional framework by [22]. Their model uses a tensor-based skipgram with negative sampling, where adjectives are transformation matrices that modify noun embeddings to capture their compositional influence in phrases. Nouns are represented as vectors, and adjectives as linear transformation matrices within the noun vector space. The model maximizes the likelihood of observing a noun in the context of a given adjective by optimizing the similarity between the adjective-transformed noun and its surrounding context words. This objective is given by the following equation:

$$\sum_{\boldsymbol{c}_t' \in \mathcal{C}_t} \log \sigma\left(\mathbf{A}\boldsymbol{n}_t \cdot \boldsymbol{c}_t'\right) + \sum_{\overline{\boldsymbol{c}_t}' \in \overline{\mathcal{C}_t}} \log \sigma\left(-\mathbf{A}\boldsymbol{n}_t \cdot \overline{\boldsymbol{c}_t}'\right) \tag{A.1}$$

where $\mathbf{A}$ is the transformation matrix representing the adjective, and $\boldsymbol{n}_t$ is the noun embedding. $\mathcal{C}_t$ and $\overline{\mathcal{C}_t}$ denote the sets of positive and negative contexts, respectively, derived from textual representations. The noun embeddings are learned using the following objective function:

$$\sum_{\boldsymbol{c}_t' \in \mathcal{C}_t} \log \sigma\left(\boldsymbol{n}_t \cdot \boldsymbol{c}_t'\right) + \sum_{\overline{\boldsymbol{c}_t}' \in \overline{\mathcal{C}_t}} \log \sigma\left(-\boldsymbol{n}_t \cdot \overline{\boldsymbol{c}_t}'\right) \tag{A.2}$$

where $\boldsymbol{n}_t$ represents the noun embedding being learned, $\mathcal{C}_t$ is the set of positive contexts, and $\overline{\mathcal{C}_t}$ denotes the negative context samples. This function maximizes the likelihood of observing positive context words near the noun while minimizing the likelihood of

negative context words, effectively capturing the contextual relationships within the noun's embedding space.

Dataset: For initial experimentation, we chose the Text8 dataset, a compact yet representative subset of English Wikipedia consisting of 17 million words. Text8 is popular for language modeling tasks and is particularly suitable for exploring phrase-level semantic relationships due to its manageable size and quality.

Implementation Details: We first preprocess the dataset by converting text to lowercase, tokenizing punctuation, and filtering out low-frequency words. A vocabulary is then created, and frequent words are subsampled to reduce noise. Using a skip-gram approach with a context window size of 5, positive context pairs are generated, with 10 negative samples drawn from a noise distribution. The model is trained with a learning rate of 0.003, optimizing noun embeddings over 10 epochs with a batch size of 512. Adjective matrices, initialized as identity, are learned by transforming noun embeddings to capture the meanings of adjective-noun phrases. For each adjective, noun pairs are extracted, and transformation matrices are trained using Cross-Entropy loss with a batch size of 32 over 10 epochs.

Results and Discussion: The evaluation process involves computing cosine similarity scores between learnt adjective-noun phrase embeddings and comparing these scores against human-annotated similarity ratings. For each adjective-noun pair, the model-generated similarity scores are matched with corresponding human judgments, and Spearman's rank correlation assesses the alignment between model-based and human perceptions.

Table A.1: Semantic and audio similarities between phrases.

	Audio Similarity	**Semantic Similarity**
TSG-Text	0.19	0.26

Results in able A.1 reveal that the text-only model, represented by *TSG-Text*, achieves a moderate correlation of 0.26 for semantic similarity and a lower correlation of 0.19 for audio similarity. This indicates that the model captures some level of semantic similarity between adjective-noun phrases but is less effective at capturing audio-based relationships, as expected given the absence of auditory context.

The higher performance on semantic similarity suggests that the text-only approach is better suited to tasks where phrases share conceptual or linguistic attributes rather than sensory characteristics. The relatively low audio similarity score highlights the limitations of a purely textual model in contexts involving perceptual qualities, such as sound. Additionally, we understand that the dataset size may not be sufficient to fully capture nuanced relationships. In future work, we plan to train the model on a larger dataset, such as the full Wikipedia corpus, to enhance its capacity for capturing detailed semantic associations.

Appendix B

A Novel Multimodal Phrase Dataset

The adjectives in the dataset are: *vocal, sad, resonant, quick, percussive, musical, melodious, melodic, mechanical, low, male, loud, instrumental, industrial, human, large, high-pitched, high, female, happy, heavy, electrical, fast, electronic, distant, digital, deep, dark, big, angry*.

B.1 Annotation Guidelines

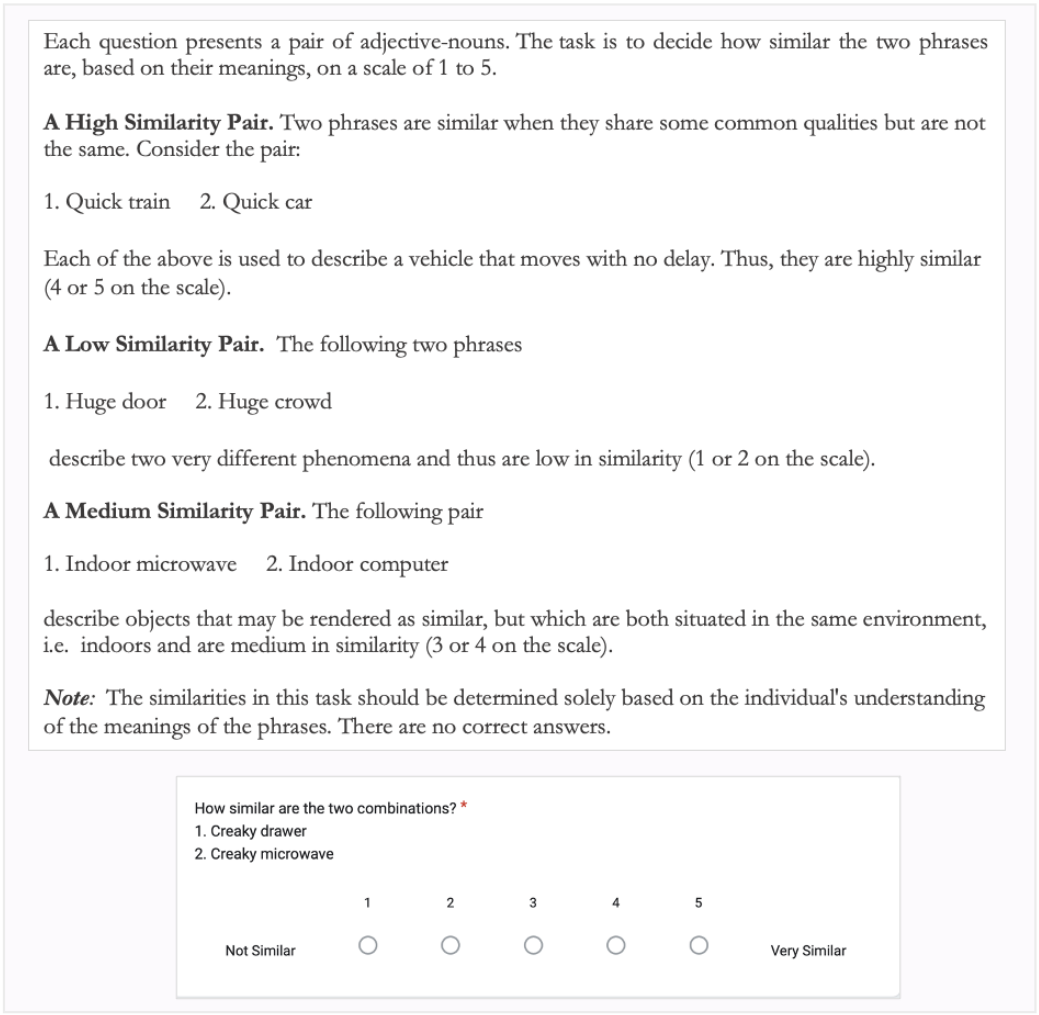
To ensure consistency across annotations in both audio and semantic similarity tasks, annotators used a 1–5 scale. A score of 1 indicated no similarity in sound or meaning, 3 represented moderate similarity with some shared qualities but notable differences, and 5 indicated high similarity, with items almost identical in sound or meaning. They were shown example pairs with sample scores to illustrate the scale but were encouraged to apply their own judgment, as interpretations could vary. The data was divided into environmental and musical categories with tailored questionnaires: environmental items focused on natural elements (e.g., *heavy rain* vs. *heavy wind*), while musical items addressed instrumental or musical qualities (e.g., *soft melody* vs. *soft harmony*).

B.1.1 Semantic Similarity

In the semantic similarity task, annotators assessed similarity based solely on meaning, ignoring any auditory components. For each pair, annotators were asked to focus on the implied qualities or conceptual overlap between phrases. Higher ratings were assigned when phrases conveyed similar ideas or emotions (e.g., *bright morning* and *bright sun*).

Examples for environmental and musical questionnaires are shown in Figures B.1 and B.2.

Figure B.1 exemplifies the semantic similarity annotation guidelines designed specifically for environmental phrases. It illustrates the structured approach annotators followed when rating phrase pairs, such as *quick train* and *quick car*, may have high similarity due to shared key attributes like speed and transportation context. Similarly Figure B.2 shows an example of semantic similarity annotation guidelines for musical phrases. Annotators were guided to assess pairs like *melodic tune* and *melodic harmony* as highly similar. For both environmental and musical categories, the guidelines differ only in the examples and their descriptions.


Each question presents a pair of adjective-nouns. The task is to decide how similar the two phrases are, based on their meanings, on a scale of 1 to 5.

A High Similarity Pair. Two phrases are similar when they share some common qualities but are not the same. Consider the pair:

1. Quick train 2. Quick car

Each of the above is used to describe a vehicle that moves with no delay. Thus, they are highly similar (4 or 5 on the scale).

A Low Similarity Pair. The following two phrases

1. Huge door 2. Huge crowd

describe two very different phenomena and thus are low in similarity (1 or 2 on the scale).

A Medium Similarity Pair. The following pair

1. Indoor microwave 2. Indoor computer

describe objects that may be rendered as similar, but which are both situated in the same environment, i.e. indoors and are medium in similarity (3 or 4 on the scale).

Note: The similarities in this task should be determined solely based on the individual's understanding of the meanings of the phrases. There are no correct answers.

Figure B.1: Example of semantic similarity annotation guidelines for environmental phrases.

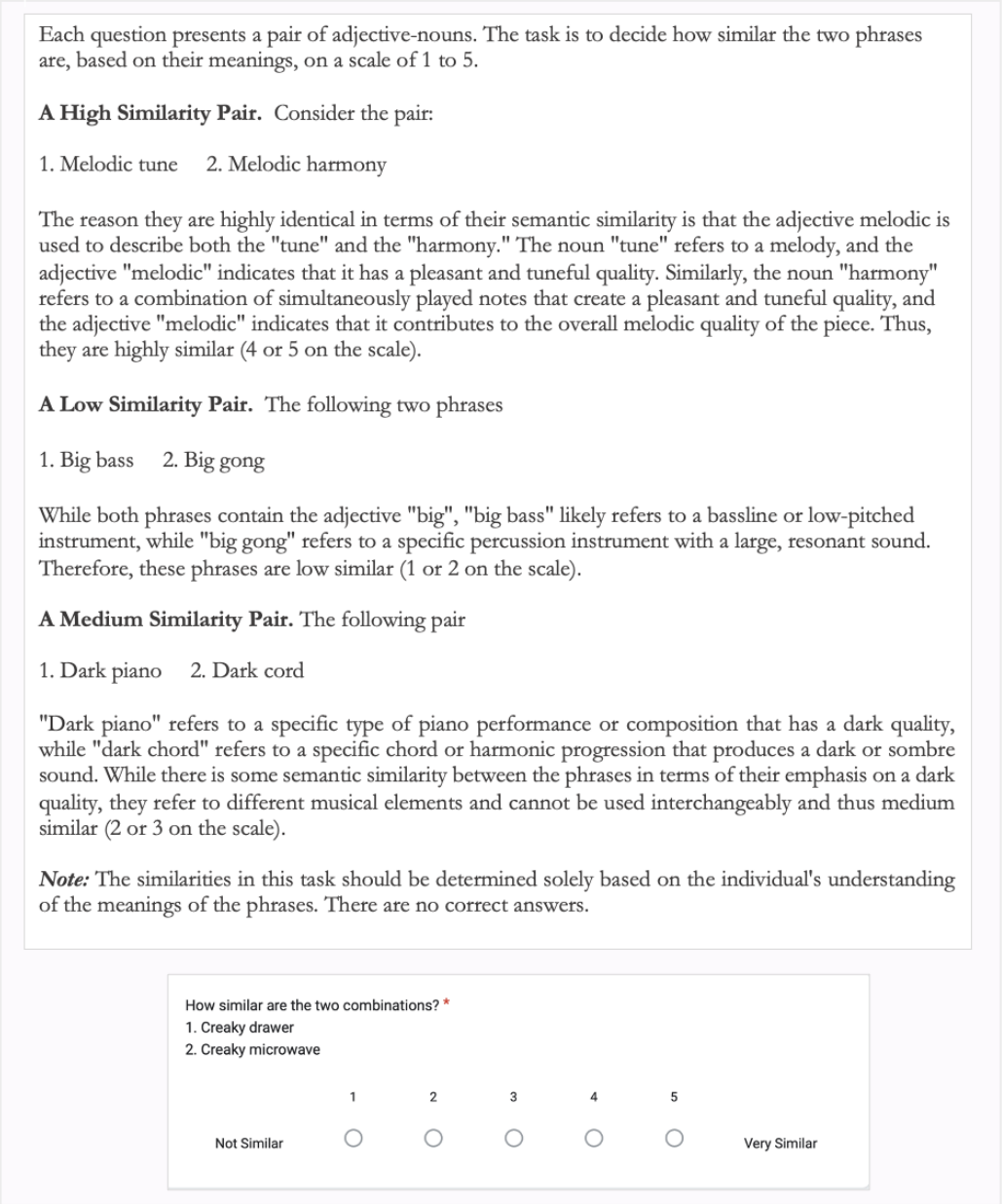


Figure B.2: Example of semantic similarity annotation guidelines for musical phrases.

B.1.2 Audio Similarity

In the audio similarity task, annotators assessed pairs of phrases based on their perception of how the sounds described by each phrase might compare to one another. They were asked to imagine the sound of each phrase and rate its similarity to the other. For instance, they might consider the sound of *heavy rain* and how similar it could be to *heavy thunderstorm*.

Annotators used the same 1-5 rating scale as in the semantic similarity task, where they rated the sound similarity between two phrases. Example pairs were provided to help clarify the concept; however, these examples were meant as guidelines and not definitive, as individual perceptions of sounds may vary. Figures B.3 and B.4 illustrate examples of the audio similarity annotation guidelines for environmental and musical sounds, respectively.

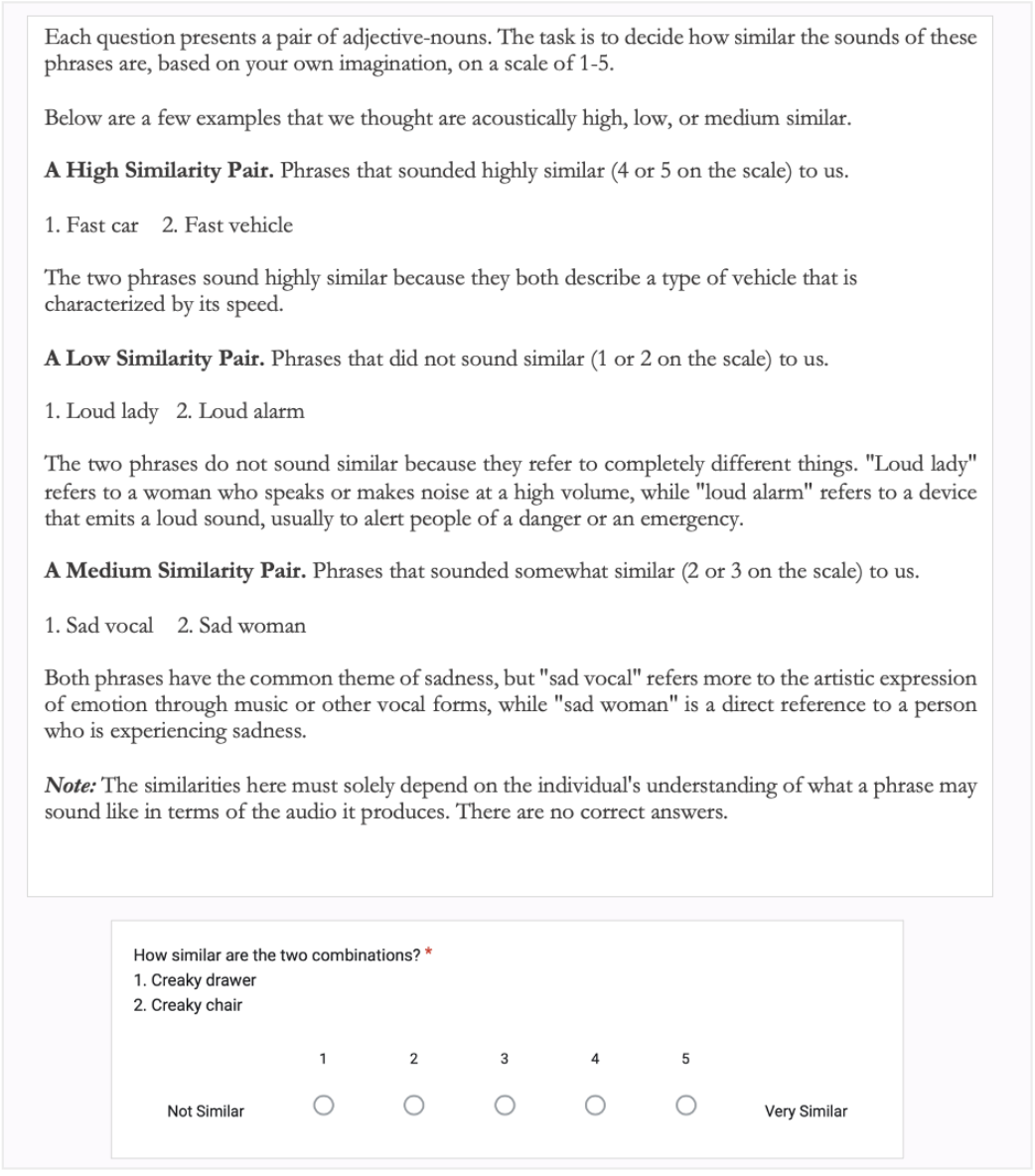


Figure B.3: Example of audio similarity annotation guidelines for environmental sounds.

In Figure B.3, an example of audio similarity guidelines for environmental sounds is presented. This figure shows how annotators were instructed to rate audio pairs like

fast car and *fast vehicle* as highly similar, with additional guidance on distinguishing medium and low similarity pairs, such as *loud lady* vs. *loud alarm*. Similarly, Figure B.4 provides an example of audio similarity guidelines for musical sounds, illustrating how pairs like *soft music* and *soft melody* are rated as highly similar and explaining how to approach medium similarity cases, such as *sad flute* vs. *sad saxophone*.

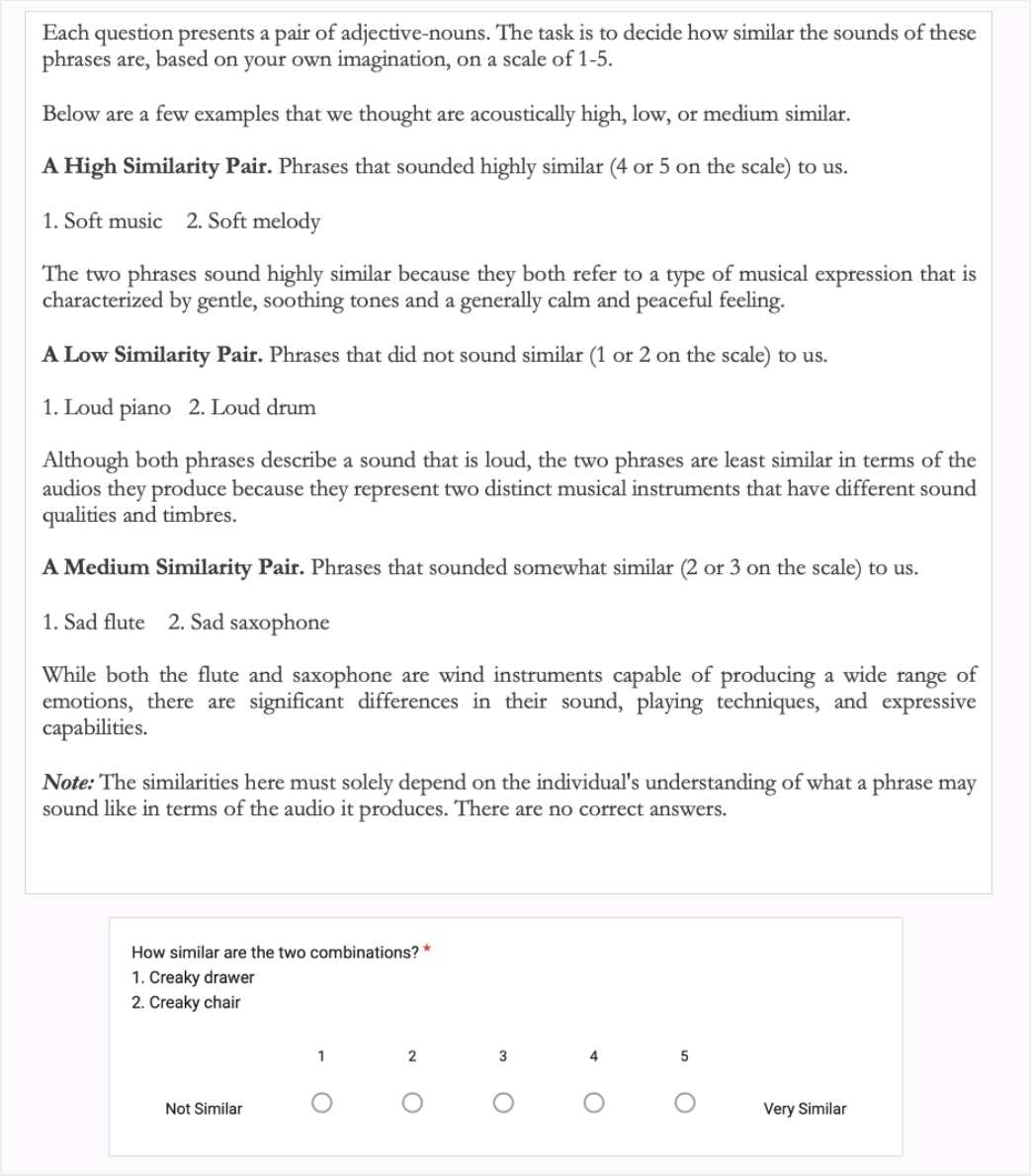


Figure B.4: Example of audio similarity annotation guidelines for musical sounds.

B.2 Data Insights

We collected human similarity ratings for 3144 adjective-noun phrase pairs across two modalities: semantic and auditory. Each pair was rated on a scale from 1 (low similarity) to 5 (high similarity) by multiple annotators. The final similarity score for each pair in both modalities is computed as the mean of all individual annotations.

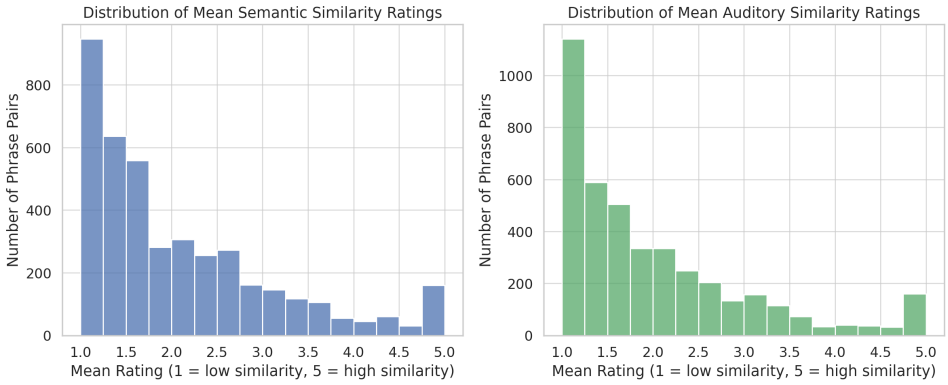


Figure B.5: Histogram of mean similarity ratings across all phrase pairs. Left: semantic similarity. Right: auditory similarity. Each bin represents a range of mean ratings (e.g., 2.0–2.25), computed by averaging multiple annotator scores per pair.

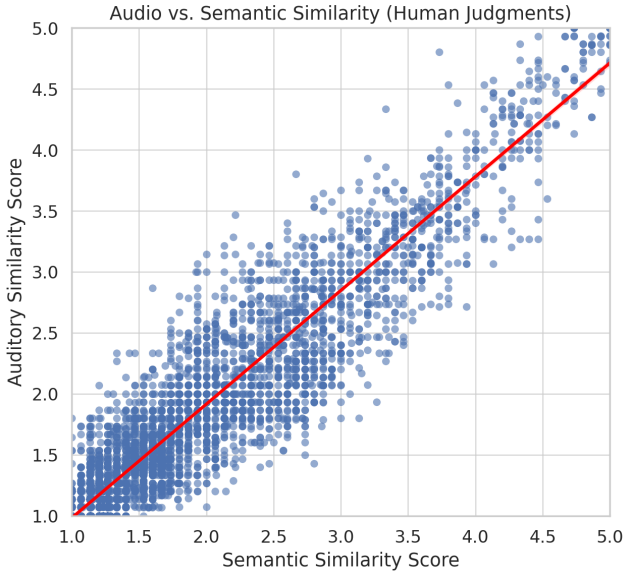


Figure B.6: Scatter plot comparing human-annotated semantic and auditory similarity scores for phrase pairs. Each point represents one phrase pair. A red diagonal line indicates the $x = y$ reference, where semantic and auditory judgments align perfectly.

Figure B.5 shows the distributions of the mean similarity ratings across phrase pairs. The

x-axis represents the average similarity score of a phrase pair, while the y-axis shows how many pairs received a score within each bin. Most ratings fall between 1.5 and 3.5, with peaks near the lower end of the scale in both modalities. This suggests that most phrase pairs were judged to be only moderately similar in either meaning or sound, while highly similar or dissimilar pairs were less frequent. The similarity in shape between the two distributions supports a strong correspondence between human perceptions of semantic and auditory similarity.

Figure B.6 shows the relationship between the mean semantic and auditory similarity ratings, both ranging from 1 (low similarity) to 5 (high similarity). Each point in the plot represents a phrase pair, with its x-coordinate corresponding to the mean semantic similarity score and its y-coordinate representing the mean auditory similarity score. The strong upward trend indicates that phrases rated as semantically similar were also likely to be judged as auditorily similar. This relationship is quantified by a Spearman correlation of 0.90, demonstrating a strong and consistent agreement between semantic and auditory similarity judgments.

Appendix C

Evaluation of the Framework

C.1 Adjective Similarities

In addition to phrase similarity experiments, we conducted a small-scale evaluation of adjective–adjective semantic similarities using the SimLex-999 dataset [9].

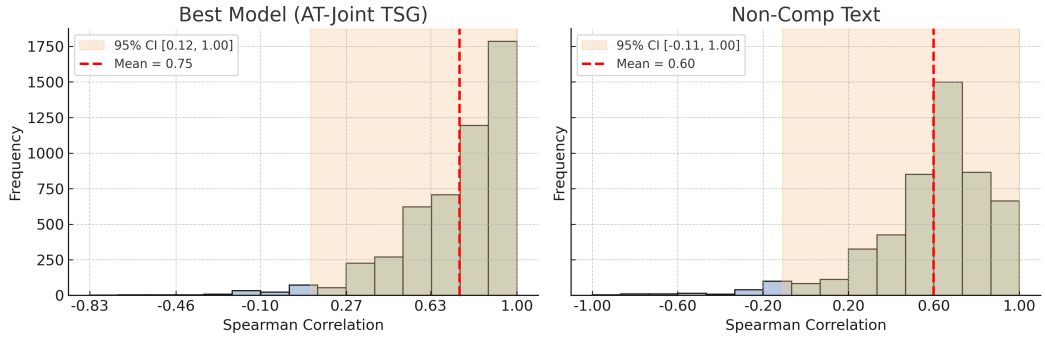


Figure C.1: Bootstrap distributions of Spearman correlations between model-predicted and human-rated adjective–adjective similarities for the best-performing model (AT-Joint TSG; left) and a unimodal baseline (Non-Comp Text; right). Distributions are based on 5,000 resamples of the evaluation pairs, with shaded 95% confidence intervals and dashed red lines showing bootstrap means.

From the audio dataset, 11 adjectives overlapped with SimLex, forming 8 evaluation pairs. For each model, we computed cosine similarities for these pairs and measured Spearman's rank correlation (ρ) with the corresponding SimLex human scores..

Given the very limited number of evaluation pairs, the results should be interpreted as indicative rather than definitive. To quantify the uncertainty associated with such a small sample, we applied a pair-level bootstrap to only the best-performing model (AT-Joint

TSG) and one unimodal baseline (Non-Comp Text) for comparison (Figure C.1). In each of 5,000 iterations, we resampled the adjective pairs with replacement, recomputed (ρ), and built a distribution of bootstrapped correlations. This approach does not alter the model or human scores but estimates how much the correlation might vary if a different set of adjective pairs were drawn from the same population.

As expected, the resulting confidence intervals are relatively wide, reflecting the small sample size and limited coverage. Nevertheless, this experiment is valuable because it links our models' performance to an established external benchmark, showing that the trends observed in the phrase similarity experiments, such as the advantage of multimodal models over unimodal ones, are also apparent, though less pronounced, at the single-word (adjective) level.

C.2 Phrase Similarities

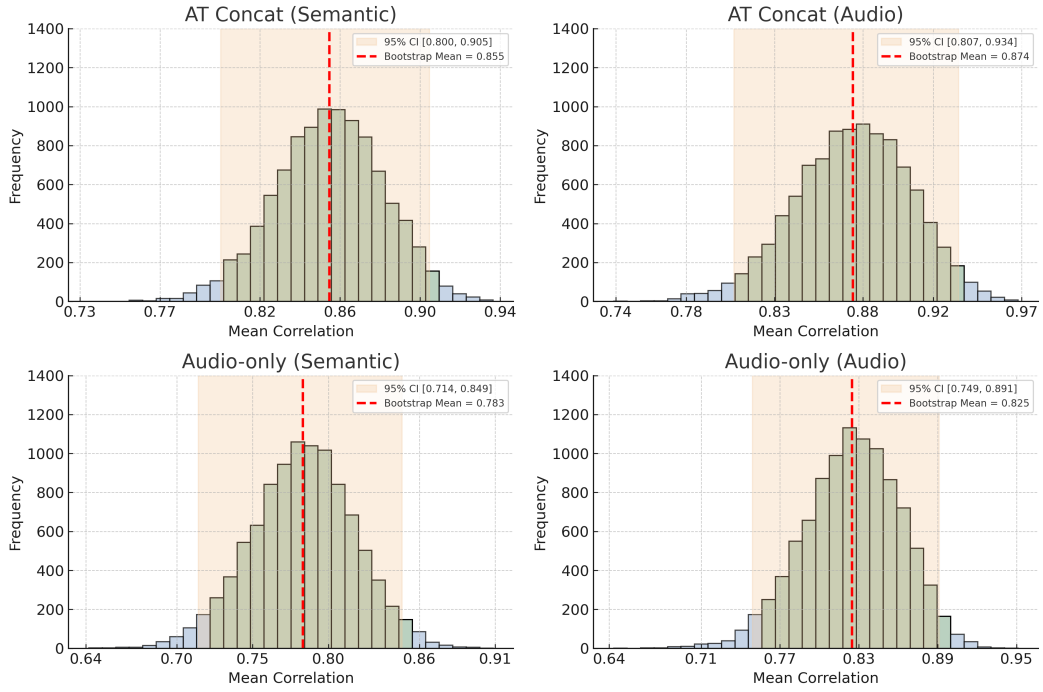


Figure C.2: Adjective-level bootstrap distributions (10,000 iterations) for four models. Left: semantic models; right: auditory models. Shaded areas show 95% CIs, red dashed lines mark bootstrap means, and a fixed y-axis scale enables direct visual comparison.

Figure C.2 presents the adjective-level bootstrap distributions of mean Spearman correlations between model predictions and human similarity judgements for four systems: AT

Concat (Semantic), AT Concat (Audio), Audio-only (Semantic), and Audio-only (Audio). Semantic models are shown in the left column and auditory models in the right column for direct comparison. Each distribution was generated using 10,000 bootstrap iterations, resampling the 30 per-adjective correlation values with replacement and computing the mean for each resample. The AT Concat (Audio) model achieved a raw mean of 0.8746 with a 95% confidence interval (CI) of [0.8070, 0.9342], while AT Concat (Semantic) reached 0.8548 (CI: [0.8003, 0.9046]). The Audio-only (Semantic) model produced a mean of 0.7832 (CI: [0.7142, 0.8486]), and Audio-only (Audio) scored 0.8248 (CI: [0.7490, 0.8911]). The bootstrap means closely match the raw means, as expected for the mean statistic, while the CI widths quantify variability due to the finite set of adjectives. Overall, the multimodal AT Concat models outperform their unimodal counterparts, with smaller CI ranges indicating more consistent performance across adjectives.

Appendix D

Application: Multimodal Recommendations

D.1 Examples

This section presents examples from the 145-program BBC dataset obtained from the suggested recommender. To interpret the final recommendations and understand the impact of individual recommenders, examples from genre, audio, and tag-based recommenders are also discussed. Table D.1 highlights two high-ranking and two low-ranking recommendations for the program EastEnders, illustrating the strengths and limitations of each recommender.

Table D.1: Recommendation examples for EastEnders

Example (Good/Bad)	User-Based Recommendations	Text-Based Recommendations		Audio-Based Recommendations	Late Fusion
		Genres	Tags		Audio +Tags + Genre
Good	Waterloo Road	Waterloo Road	Sexy Beasts	Waterloo Road	Doctors
Good	Outnumbered	Death in Paradise	Live at the Electric	The Notorious Bettie Page	Waterloo Road
Bad	Meet the Author	Dissected	Salamander	Mastermind	The Football League Show
Bad	Italy Unpacked	Top of the Pops	The Football League Show	Who Dares Wins	University Challenge

Table D.2 presents metadata for each episode, offering insight into the behavior of each recommender. For audios, first 10 audio words with highest counts based on audio histograms are used for testing. Since audio words cannot be directly interpreted like text, authors' performed a manual classification of the audio data. Each audio sample

was categorized according to whether its music and voiceover fit a soft, moderate, or loud audio profile.

Table D.2: Metadata of examples, including ID, Title, Genre, Tags, and Audio for each program. Only the top 10 highest-weighted starfruit tags are displayed.

ID	Title	Genre	Mention Tags (10/episode)	Audio Words (10/episode)
b03vznpt	EastEnders	drama/soaps	dogs, family, Cardiff, cheating, community, Brighton, love, daughter, children, protest	Music: soft, Voiceover: soft Audio words: 19, 44, 0, 7, 46, 31, 33, 3, 21, 4,
b03w0d8z	Waterloo Road	drama	Campaigning, love, eating, community, cheating, office, future, exams, daughter, crime	Music: soft , Voiceover: soft Audio words: 19, 46,3 1, 21, 44, 0, 7, 34, 1, 17,
b03w7snk	Outnumbered	comedy/sitcoms	animals, university, eating, students, nature, drama, bullfighting, cake, rapping, romance	Music: soft, Voiceover: moderate Audio words: 19, 44, 21, 46, 0, 7, 16, 4, 5, 25,
b03w790q	Death in Paradise	drama/crime	birds, Vietnam, hobby, eating, theft, war, friendship, English, murder, law	Voice: moderate, Music: low Audio words: 31, 0, 46, 44, 7, 2, 21, 19, 10, 32
b03wcmdc	Sexy Beasts	factual/familiesandrelationships	London, dating, romance, eating, relationships, love, public, relations, future, future, cheating	Music: moderate, Voice: loud Audio words: 46, 0, 4, 31, 19, 41, 5, 25, 45, 24
b03v3n0d	Live at the Electric	comedy/standup, entertainment	Australia, comedy, family, eating, France, future, England, love, philosophy, gender	Music: moderate ,Voice: loud Audio words: 37, 23, 31, 44, 4, 18, 1, 0, 26, 6
b00nx10r	The Notorious Bettie Page	drama/biographical	music, entertainment, film, children, dance, love, Tennessee, literature, shame, Hollywood	Music: soft , Voice: soft Audio words: 19, 44, 21, 0, 46, 10, 31, 17, 7, 34,
b03vs7g1	Doctors	drama/medical, drama/soaps	abortion, daughter, Zara, adoption, father, strokes, sex, homophobia, children, love	Music: soft, Voiceover: soft Audio words: 19, 21, 44, 10, 31, 46, 33, 0, 35, 27,
b03w7s0n	Meet the Author	factual/artscultureandthemedia, factual/artscultureandthemedia/arts	curling, David, Ukraine, winter, love, history, Stranraer, libraries, hacking, phone	Music: low, Voiceover: Moderate Audio words: 4, 41, 10, 3, 44, 24, 46, 21, 25, 19
b03qg00y	Italy Unpacked	factual/artscultureandthemedia/arts	food, Italy, Rome, Europe, cookery, sculpture, Egypt, garden, art, music	Music: Moderate, Voiceover: Moderate Audio words: 2, 0, 46, 44, 21, 42, 10, 19, 30, 32
p01mv2md	Dissected	factual	biology, animals, Charles, human, ethics, scientist, engineering, language	Music: low. Voiceover: low Audio words: 19, 21, 0, 44, 46, 10, 45, 3, 22, 26
b03mpphw	Top of the Pops	factual/artscultureandthemedia/arts, music/classicpopandrock	music, Nile, Rodgers, northern, Ireland, Coventry, pop, disco, British, army	Music: Loud, Voiceover: Moderate Audio words: 17, 34, 46, 31, 0, 21, 37, 44, 19, 45
b01pyjxw	Salamander	drama/crime	idles	Music: low, Voiceover:Moderate Audio words: 7, 1, 21, 31, 0, 44, 29, 19, 34, 18
b03wc7gf	The Football League Show	sport/football	football, championship, city Birmingham, town, Huddersfield, Vale, port, Brentford, Millwall	Music: low, Voiceover: high Audio words: 4, 21, 0, 49, 13, 17, 36, 31, 45, 42
b03wby67	Mastermind	entertainment	London, Nottingham, history, criticism, Caerphilly, war, tiring, tennis, Belize, arts	Music: low, Voiceover: high Audio words: 44, 23, 2, 26, 36, 5, 43, 48, 12, 47
b03w4c7c	Who Dares Wins	entertainment	television, Cuba, Abba, Jamaica, Kidderminster, national, Mayall, Rik, hull, DC	Music: Moderate, Voiceover: high Audio words: 37, 19, 3, 0, 1, 42, 13, 36, 45, 5
b03w7vbz	University Challenge	entertainment	Cardiff, science, Faso, Burkina, Benin, history, criticism, linguistics, Huntingdon, physics	Music: low, Voiceover: Moderate Audio words: 44, 2, 23, 47, 13, 43, 26, 28, 49, 36

From the final late fusion model for *EastEnders*, a good example is *Waterloo Road*, which aligns well due to its shared genre of drama and themes around community and social challenges. This thematic similarity is enhanced by soft audio elements and matching tags like *love* and *family*, reflecting the relational aspects that connect well with *EastEnders* fans. In contrast, a poor recommendation example is *The Football League Show*. Although it may share a broad audience base with *EastEnders*, its focus on sports events, paired with louder audio profiles and tags unrelated to *EastEnders*' community-oriented themes, makes it less compatible. This illustrates how genre, thematic alignment, and audio profile differences can affect recommendation quality.

Overall, this approach enhances the recommender system's ability to align suggestions with user preferences by integrating both content and audio characteristics, creating a more personalized user experience.

Bibliography

[1] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[2] John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

[3] Anne Harris and Stacy Holman Jones. Words. In *Writing for Performance*, pages 19–35. Brill Sense, 2016.

[4] Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[5] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[6] G. Corrado T. Mikolov, K. Chen and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, AZ, 2013.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

[10] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414, 2001.

[11] N. Tran E. Bruni and M. Baroni. Multimodal distributional semantics. *Journal of Artifical Intelligence Research*, 49:1–47, 2014.

[12] Richard Montague et al. *English as a formal language*. Ed. di Comunità, 1970.

[13] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244, 2008.

[14] William Blacoe and Mirella Lapata. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 546–556, 2012.

[15] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211, 2012.

[16] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017.

[17] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[18] Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216, 1990.

[19] Stephen Clark and Stephen Pulman. Combining symbolic and distributional models of meaning. 2007.

[20] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*, 2010.

[21] Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1183–1193, 2010.

[22] Jean Maillard and Stephen Clark. Learning adjective meanings with a tensor-based skip-gram model. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 327–331, 2015.

[23] Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. Multi-step regression learning for compositional distributional semantics. *arXiv preprint arXiv:1301.6939*, 2013.

[24] Gijs Wijnholds, Mehrnoosh Sadrzadeh, and Stephen Clark. Representation learning for type-driven composition. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 313–324, Online, November 2020. Association for Computational Linguistics.

[25] Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver. Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 708–719, 2014.

[26] Arthur M Glenberg and David A Robertson. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, 43(3):379–401, 2000.

[27] Yansong Feng and Mirella Lapata. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 91–99. Association for Computational Linguistics, 2010.

[28] Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, 2014.

[29] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47, 2014.

[30] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015.

[31] Douwe Kiela and Stephen Clark. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, 2015.

[32] Alessandro Lopopolo and Emiel van Miltenburg. Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 70–75, 2015.

[33] Douwe Kiela and Stephen Clark. Learning neural audio embeddings for grounding semantics in auditory perception. *Journal of Artificial Intelligence Research*, 60:1003–1030, 2017.

[34] Wenhao Zhu, Xiaping Xu, Ke Yan, Shuang Liu, and Xiaoya Yin. A synchronized word representation method with dual perceptual information. *IEEE Access*, 8:22335–22344, 2020.

[35] Saba Nazir, Taner Cagali, Mehrnoosh Sadrzadeh, and Chris Newell. Audiovisual, genre, neural and topical textual embeddings for tv programme content representation. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 197–200, 2020.

[36] Guy Emerson. What are the goals of distributional semantics? *arXiv preprint arXiv:2005.02982*, 2020.

[37] Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*. Association for Computational Linguistics, 2024.

[38] Hadi Wazni, Kin Lo, and Mehrnoosh Sadrzadeh. Verbclip: Improving verb understanding in vision-language models with compositional structures. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 195–201, 2024.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[40] Jean Maillard, Stephen Clark, and Edward Grefenstette. A type-driven tensor-based semantics for ccg. In *Proceedings of the EACL 2014 Type Theory and Natural Language Semantics Workshop*, pages 46–54, 2014.

[41] Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

[42] Eva M Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive science*, 41(1):102–136, 2017.

[43] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.

[44] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020.

[45] Karo Moilanen and Stephen Pulman. Sentiment composition. In *International Conference Recent Advances in Natural Language Processing, RANLP*. ACL Anthology, 2007.

[46] Ainur Yessenalina and Claire Cardie. Compositional matrix-space models for sentiment analysis. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 172–182, 2011.

[47] Shima Asaadi and Sebastian Rudolph. Gradual learning of matrix-space models of language for sentiment analysis. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 178–185, 2017.

[48] Tao Chen, Felix X Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. Object-based visual sentiment concept analysis and application. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 367–376, 2014.

[49] Zuhe Li, Qian Sun, Qingbing Guo, Huaiguang Wu, Lujuan Deng, Qiuwen Zhang, Jianwei Zhang, Huanlong Zhang, and Yu Chen. Visual sentiment analysis based on image caption and adjective–noun–pair description. *Soft Computing*, pages 1–13, 2021.

[50] Bo Yang, Tao Mei, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, and Mingjing Li. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, page 73–80, New York, NY, USA, 2007. Association for Computing Machinery.

[51] Hazım Kemal Ekenel and Tomas Semela. Multimodal genre classification of tv programs and youtube videos. *Multimedia tools and applications*, 63(2):547–567, 2013.

[52] Konstantinos Bougiatiotis and Theodoros Giannakopoulos. Enhanced movie content similarity based on textual, auditory and visual information. *Expert Systems with Applications*, 96:86 – 102, 2018.

[53] Saba Nazir and Mehrnoosh Sadrzadeh. How does an adjective sound like? exploring audio phrase composition with textual embeddings. In *Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning*, pages 13–18, 2024.

[54] Saba Nazir and Mehrnoosh Sadrzadeh. The potential of multimodal compositionality for enhanced recommendations through sentiment analysis. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, pages 26–30, 2024.

[55] Saba Nazir, Taner Cagali, Chris Newell, and Mehrnoosh Sadrzadeh. Cosine similarity of multimodal content vectors for tv programmes. In *Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020*, 2020.

[56] Saba Nazir, Taner Cagali, Mehrnoosh Sadrzadeh, and Chris Newell. Audiovisual, genre, neural and topical textual embeddings for tv programme content representation. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 197–200. IEEE, 2020.

[57] Taner Cagali, Hadi Wazni, Saba Nazir, Mehrnoosh Sadrzadeh, and Chris Newell. Semantic and lexical token based vectors improve precision of recommendations for tv programmes. In *2023 IEEE International Symposium on Multimedia (ISM)*, pages 287–290. IEEE, 2023.

[58] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208, 1996.

[59] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

[60] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[61] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[62] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[63] David R. Dowty, Robert E. Wall, and Stanley Peters. *Introduction to Montague Semantics*. Dordrecht, 1981.

[64]

[65] Tony Plate et al. Holographic reduced representations: Convolution algebra for compositional distributed representations. In *IJCAI*, pages 30–35, 1991.

[66] Katrin Erk and Sebastian Padó. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, 2008.

[67] Emiliano Raul Guevara. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 workshop on geometrical models of natural language semantics*, pages 33–37, 2010.

[68] Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman. Concrete sentence spaces for compositional distributional models of meaning. In *Proceedings of IWCS*, pages 125–134, 2011.

[69] Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimental support for a categorical compositional distributional model of meaning. In Regina Barzilay and Mark Johnson, editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[70] Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimenting with transitive verbs in a discocat. *arXiv preprint arXiv:1107.3119*, 2011.

[71] Gijs Wijnholds and Mehrnoosh Sadrzadeh. Evaluating composition models for verb phrase elliptical sentence embeddings. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 261–271, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[72] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174, 2018.

[73] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.

[74] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[75] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[76]

[77] Khaled Koutini, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2753–2757. ISCA, 2022.

[78] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[79] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[80] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

[81] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[82] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[83] Fujishima Takuya. Realtime chord recognition of musical sound: Asystem using common lisp music. In *Proceedings of the International Computer Music Conference 1999, Beijing*, 1999.

[84] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968. IEEE, 2014.

[85] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2015.

[86] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

[87] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[88] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[89] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

[90] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[91] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[92] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A

large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[93] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[94] Anurag Kumar, Maksim Khadkevich, and Christian Fügen. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 326–330. IEEE, 2018.

[95] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856. IEEE, 2019.

[96] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[97] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Bjoern W Schuller. Audio self-supervised learning: A survey. *Patterns*, 3(12), 2022.

[98] Yu-An Chung and James Glass. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*, 2018.

[99] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[100] Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350. IEEE, 2021.

[101] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.

[102] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[103] Joachim Lambek. Type grammar revisited. In *Logical Aspects of Computational Linguistics: Second International Conference, LACL'97 Nancy, France, September 22-24, 1997 Selected Papers 2*, pages 1–27. Springer, 1999.

[104] Mark Steedman. *The syntactic process*. MIT press, 2001.

[105] Mark Steedman. Mark steedman, the syntactic process (language, speech, and communication). cambridge, ma: Mit press, 2000. pp. xiv 330. *Journal of Linguistics*, 38(3):645–708, 2002.

[106] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[107] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[108] Stephen Clark and James R. Curran. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552, 2007.

[109] Julia Hockenmaier and Mark Steedman. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn treebank. *Computational Linguistics*, 33(3):355–396, 2007.

[110] Ron Wehrens and B-H Mevik. The pls package: principal component and partial least squares regression in r. 2007.

[111] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, Silvia Bernardini, et al. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, 2008.

[112] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[113] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

[114] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

[115] Amitava Das and Sivaji Bandyopadhyay. Opinion-polarity identification in bengali. In *International conference on computer processing of oriental languages*, pages 169–182. Chinese and Oriental Languages Computer Society California, USA, 2010.

[116] Kishorjit Nongmeikapam, Dilipkumar Khangembam, Wangkheimayum Hemkumar, Shinghajit Khuraijam, and Sivaji Bandyopadhyay. Verb based manipuri sentiment analysis. *Int J Nat Lang Comput*, 3(3):113–118, 2014.

[117] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[118] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[119] T Wilson. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP*, 2005.

[120] Wanxiang Che, Yanyan Zhao, Honglei Guo, Zhong Su, and Ting Liu. Sentence compression for aspect-based sentiment analysis. *IEEE/ACM Transactions on audio, speech, and language processing*, 23(12):2111–2124, 2015.

[121] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615, 2016.

[122] Alexander Hogenboom, Bas Heerschop, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. *Decision support systems*, 62:43–53, 2014.

[123] Shahla Nemati and Ahmad Reza Naghsh-Nilchi. Exploiting evidential theory in the fusion of textual, audio, and visual modalities for affective music video retrieval. In *2017 3rd international conference on pattern recognition and image analysis (ipria)*, pages 222–228. IEEE, 2017.

[124] Alvaro Ortigosa, José M Martín, and Rosa M Carro. Sentiment analysis in facebook and its application to e-learning. *Computers in human behavior*, 31:527–541, 2014.

[125] Yashpalsing Chavhan, ML Dhore, and Pallavi Yesaware. Speech emotion recognition using support vector machine. *International Journal of Computer Applications*, 1(20):6–9, 2010.

[126] Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. Multimodal sentiment analysis of spanish online videos. *IEEE intelligent systems*, 28(3):38–45, 2013.

[127] Nadia FF Da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. Tweet sentiment analysis with classifier ensembles. *Decision support systems*, 66:170–179, 2014.

[128] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.

[129] Rakhee Sharma, Ngoc Le Tan, and Fatiha Sadat. Multimodal sentiment analysis using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1475–1478. IEEE, 2018.

[130] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 163–171, 2017.

[131] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn Schuller, and Kurt Keutzer. Affective image content analysis: A comprehensive survey. 2018.

[132] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. Survey on visual sentiment analysis. *IET Image Processing*, 14(8):1440–1456, 2020.

[133] Ringki Das and Thoudam Doren Singh. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38, 2023.

[134] Upendra Singh, Kumar Abhishek, and Hiteshwar Kumar Azad. A survey of cutting-edge multimodal sentiment analysis. *ACM Computing Surveys*, 56(9):1–38, 2024.

[135] Sebastian Rudolph and Eugenie Giesbrecht. Compositional matrix-space models of language. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 907–916, 2010.

[136] Svetlana Kiritchenko and Saif M. Mohammad. The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of NAACL-HLT 2016*, pages 43–52, San Diego, California, June 2016. Association for Computational Linguistics.

[137] Svetlana Kiritchenko and Saif M. Mohammad. Sentiment composition of words with opposing polarities. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1102–1108, San Diego, California, June 2016. Association for Computational Linguistics.

[138] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232, 2013.

[139] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In

Proceedings of the 45th annual meeting of the association of computational linguistics, pages 440–447, 2007.

[140] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[141] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 70–74, 2007.

[142] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[143] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[144] Raymond J Mooney and Loriene Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204, 2000.

[145] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12:331–370, 2002.

[146] T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[147] P. Turney and P. Pantel. From frequency to meaning: vector space models of semantics. *Journal of Artifical Intelligence Research*, 37:141–188, 2010.

[148] Qiang Zhu, Mei-Chen Yeh, and Kwang-Ting Cheng. Multimodal fusion using learned text concepts for image categorization. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 211–220, 2006.

[149] Oren Barkan, Noam Koenigstein, Eylon Yogev, and Ori Katz. Cb2cf: a neural multiview content-to-collaborative filtering model for completely cold item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 228–236, 2019.

[150] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[151] Christos H Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.

[152] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[153] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[154] Yang Liu, Wan-Lei Zhao, Chong-Wah Ngo, Chang-Sheng Xu, and Han-Qing 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*, pages 89–96, 2010.

[155] Stephanie Pancoast and Murat Akbacak. Bag-of-audio-words approach for multimedia event classification. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[156] Shourabh Rawat, Peter F Schulam, Susanne Burger, Duo Ding, Yipei Wang, and Florian Metze. Robust audio-codebooks for large-scale event detection in consumer videos. In *INTERSPEECH*, pages 2929–2933, 2013.

[157] Axel Plinge, Rene Grzeszick, and Gernot A Fink. A bag-of-features approach to acoustic event detection. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3704–3708. IEEE, 2014.

[158] Rene Grzeszick, Axel Plinge, and Gernot A Fink. Temporal acoustic words for online acoustic event detection. In *German Conference on Pattern Recognition*, pages 142–153. Springer, 2015.

[159] Hyungjun Lim, Myung Jong Kim, and Hoirin Kim. Robust sound event classification using lbp-hog based bag-of-audio-words feature representation. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[160] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[161] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[162] Ebrahim Karami, Siva Prasad, and Mohamed Shehata. Image matching using sift, surf, brief and orb: performance comparison for distorted images. *arXiv preprint arXiv:1710.02726*, 2017.

[163] MyMediaLite. Mymedialite recommender system library. `http://www.mymedialite.net/`. (Accessed on 02/18/2020).

[164] Ayşegül Özkaya Eren and Mustafa Sert. Automated audio captioning using audio event clues. *arXiv preprint arXiv:2204.08567*, 2022.