

Interpreting Gender-Related Vocal Diversity through wav2vec 2.0 Speech Embedding Analysis

Ber dos Santos Neves, Ana Carolina Constantini and Romis Attux

Abstract—Self-supervised speech models such as wav2vec 2.0 produce high-dimensional embeddings that encode multiple aspects of vocal production, although the internal organization of these representations remains difficult to interpret. This work investigates the geometric structure of speaker-level speech embeddings extracted from a fine-tuned wav2vec 2.0 model using dimensionality reduction and acoustic analysis. Speaker representations were obtained from the Mozilla Common Voice 14.0 dataset and restricted to speakers in the *twenties* and *thirties* age groups in order to reduce age-related variability in vocal production. Embeddings were projected using Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP), and the resulting spaces were analyzed visually and quantitatively through the Davies–Bouldin index. Additionally, the projections were compared with the median fundamental frequency (F_0) of each speaker to investigate the relationship between latent organization and interpretable acoustic information. The results show that both PCA and UMAP preserve meaningful acoustic structure, revealing gradual F_0 transitions across the latent space. PCA produced more stable global organization, while UMAP emphasized local neighborhood structures. Cluster separability improved in binary classification settings, particularly in the *male* versus *female* configuration, while the three-class scenario exhibited stronger overlap. The analysis also indicated that the *other* category appeared consistently closer to the *female* class than to the *male* class, suggesting asymmetric acoustic relationships between vocal categories. These findings contribute to a more interpretable understanding of self-supervised speech representations through the combined analysis of dimensionality reduction, acoustic descriptors, and latent space geometry.

Keywords—Speech embeddings, gender studies, wav2vec 2.0, Transformers, dimensionality reduction, speech representation learning, machine listening, acoustic analysis.

I. INTRODUCTION

Recent advances in self-supervised learning have substantially expanded the use of deep neural networks in speech analysis. Models such as wav2vec 2.0 learn high-dimensional representations directly from raw audio signals and achieve strong performance across multiple speech-related tasks [1]. Within the classification process, these models produce latent representations that encode multiple aspects of vocal production, including phonetic content, speaker-related information, and vocal acoustic properties.

Ber dos Santos Neves, School of Electrical and Computer Engineering (FEEC), University of Campinas (UNICAMP), Campinas-SP, Brazil, e-mail: belaber.neves@gmail.com; Ana Carolina Constantini, Faculty of Medical Sciences (FCM), University of Campinas (UNICAMP), Campinas-SP, Brazil, e-mail: ana287@unicamp.br; Romis Attux, School of Electrical and Computer Engineering (FEEC), University of Campinas (UNICAMP), Campinas-SP, Brazil, e-mail: attux@unicamp.br. This work was partially supported by CAPES and CNPq (308425/2023-5).

As argued by Schlichter [2], vocalicity occupies a paradoxical position between material embodiment and social intelligibility, making the voice a particularly relevant object for studies investigating how machine learning representations encode socially meaningful acoustic structures. From this perspective, speech embeddings may be understood not only as computational representations of speech signals, but also as latent spaces that potentially encode acoustically and socially mediated vocal characteristics.

Although these embeddings are highly effective for downstream tasks, their internal organization remains difficult to interpret. Traditional speech analysis methods rely on hand-crafted acoustic descriptors, such as fundamental frequency (F_0), jitter, shimmer, and harmonic-to-noise ratio (HNR), which provide explicit and physically interpretable measurements [3]. In contrast, deep speech embeddings are distributed representations without directly interpretable dimensions, making it difficult to understand how acoustic information is organized in the latent space.

Dimensionality reduction techniques provide an important strategy for investigating the structure of these representations. Among the most commonly used methods, Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) offer complementary perspectives on the geometry of embedding spaces. PCA performs a linear projection that preserves orthogonal directions of maximum variance and provides a global view of the data distribution [4]. UMAP, in turn, is a nonlinear method that emphasizes local neighborhood relationships and can reveal manifold-like structures not captured by linear transformations [5].

This interpretability challenge becomes particularly relevant in gender-related speech analysis. While supervised classifiers may achieve high predictive performance, classification accuracy alone does not explain how gender-associated acoustic information is spatially represented in the embedding space, nor whether such representations form clearly separated groups or continuous acoustic transitions. Recent studies suggest that self-supervised speech embeddings often exhibit gradual transitions and overlapping regions rather than sharply separated clusters [6]. Acoustic properties related to vocal pitch may emerge as continuous gradients across the latent space, indicating that these representations preserve meaningful acoustic structure even in the absence of explicit class boundaries.

Despite the increasing use of self-supervised speech models, relatively little is known about how acoustically interpretable information is organized within their latent spaces. Motivated by this gap, this work investigates the geometric organization of speaker-level speech embeddings extracted from a fine-tuned wav2vec 2.0 model. Rather than focusing on clas-

sification performance, our objective is to investigate how interpretable acoustic information is organized within speaker-level self-supervised speech embeddings.

The analysis was performed using speaker-level representations obtained from the Mozilla Common Voice 14.0 dataset. To reduce age-related variability in vocal production, the study is restricted to speakers in the *twenties* and *thirties* age groups. Speaker embeddings are projected using PCA and UMAP, and the resulting spaces are analyzed both visually and quantitatively.

In particular, we investigate the relationship between the projected embeddings and the median fundamental frequency (F_0) of each speaker, using F_0 as an interpretable acoustic reference for analyzing spatial organization in the latent space. Additionally, cluster separability is evaluated through the Davies–Bouldin index under different label configurations, including multiclass and binary settings. The results provide a comparative analysis of linear and nonlinear projections of self-supervised speech representations, contributing to a more interpretable understanding of how pitch-related acoustic information is organized in speaker-level embedding spaces.

II. METHODOLOGY

This work analyzes speaker-level speech embeddings extracted from a fine-tuned wav2vec 2.0 model [1]. The methodological pipeline comprises dataset filtering, representation learning, speaker-level aggregation, dimensionality reduction, acoustic analysis, and cluster separability evaluation.

A. Dataset and Demographic Filtering

Speech samples were obtained from the Mozilla Common Voice 14.0 dataset [7] [8]. To reduce age-related variability in vocal production, only speakers in the *twenties* and *thirties* age groups were considered.

This restriction was motivated by physiological aspects of voice production. Earlier age ranges may still present developmental changes associated with vocal tract maturation [9], while later age ranges tend to exhibit stronger aging-related effects, including changes in pitch and phonatory stability [10]. Limiting the dataset to these age groups reduces this source of variability and allows the analysis to focus more directly on the learned representations.

Table I summarizes the final speaker distribution.

TABLE I
SPEAKER DISTRIBUTION BY CLASS AND AGE GROUP.

Class	Twenties	Thirties	Total
Female	201	76	277
Male	221	94	315
Other	177	59	236
Tot al	599	229	828

The dataset was divided at the speaker level into training, validation and test partitions using a 70/15/15 proportion, avoiding speaker leakage across splits.

B. Wav2vec 2.0 Embeddings

Wav2vec 2.0 is a self-supervised speech representation model based on a Transformer architecture [1]. Given a raw waveform x , the model first applies a convolutional encoder that maps the signal into latent representations for T time steps:

$$\mathbf{Z} = f_{\text{CNN}}(x), \quad (1)$$

where $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T\}$ denotes the sequence of latent vectors.

These vectors are processed by a Transformer encoder that models temporal dependencies through self-attention:

$$\mathbf{H} = f_{\text{Transformer}}(\mathbf{Z}), \quad (2)$$

where $\mathbf{H}$ represents the contextualized embeddings.

In this work, we employed the *facebook/wav2vec2-large-960h-lv60-self* pretrained model [11], which was subsequently fine-tuned for gender-related voice classification using the labels *male*, *female*, and *other*. The fine-tuning process was evaluated through classification performance, and also, through the geometric separability of the resulting embeddings under PCA and UMAP projections.

The model was fine-tuned using the AdamW optimizer with a learning rate of 3×10^{-5} , weight decay of 0.005, batch size of 12, and 30 training epochs. Embeddings were extracted from the final hidden representations and later aggregated at the speaker level.

C. Speaker-Level Aggregation

Utterance-level embeddings were aggregated at the speaker level. Let $\mathbf{e}_{ij} \in \mathbb{R}^d$ denote the embedding of utterance j from speaker i . A single speaker representation was obtained through median aggregation:

$$\mathbf{s}_i = \text{median}(\mathbf{e}_{i1}, \mathbf{e}_{i2}, \dots, \mathbf{e}_{in_i}), \quad (3)$$

where n_i is the number of utterances associated with speaker i . The median was adopted to reduce the influence of outlier utterances.

The resulting speaker-level embeddings were then stacked row-wise to form the embedding matrix

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_N \end{bmatrix} \in \mathbb{R}^{N \times d}, \quad (4)$$

where N is the number of speakers and d is the embedding dimensionality.

Speaker labels were assigned by majority voting:

$$y_i = \arg \max_c \sum_{j=1}^{n_i} \mathbb{I}(y_{ij} = c), \quad (5)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

D. Dimensionality Reduction

To analyze the geometric organization of the embeddings, we applied Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP).

PCA is a linear projection method that identifies orthogonal directions of maximum variance. Given centered embeddings $\mathbf{S}$, the covariance matrix is defined as:

$$\mathbf{C} = \frac{1}{N-1} \mathbf{S}^T \mathbf{S}, \quad (6)$$

where N is the number of speakers. The low-dimensional projection is obtained through:

$$\mathbf{Y}_{\text{PCA}} = \mathbf{S} \mathbf{W}_2, \quad (7)$$

where $\mathbf{W}_2$ contains the principal eigenvectors of $\mathbf{C}$.

UMAP is a nonlinear dimensionality reduction method designed to preserve local neighborhood relationships in the data [5]. In this study, PCA and UMAP were used as complementary approaches for investigating the organization of speaker-level embeddings.

E. Acoustic Analysis

To relate the embeddings to interpretable acoustic information, we analyzed the median fundamental frequency (F_0) of each speaker. Speaker-level F_0 values were used to color the PCA and UMAP projections, allowing the visualization of continuous acoustic gradients across the embedding space.

The analysis uses pitch as an interpretable acoustic reference for investigating how continuous vocal properties are organized within the latent space, without assuming F_0 as an isolated or sufficient class discriminator.

F. Davies–Bouldin Index

Cluster separability was quantified using the Davies–Bouldin (DB) index, which evaluates the ratio between within-cluster dispersion and between-cluster separation. Lower values indicate more compact and better separated clusters [12].

For each class C_i , the centroid is defined as:

$$\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x}. \quad (8)$$

The within-cluster dispersion is computed as:

$$S_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2, \quad (9)$$

while the distance between cluster centroids is given by:

$$M_{ij} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2. \quad (10)$$

The similarity between clusters is defined as:

$$R_i = \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right), \quad (11)$$

and the final Davies–Bouldin index is computed as:

$$DB = \frac{1}{K} \sum_{i=1}^K R_i, \quad (12)$$

where K is the number of clusters.

The DB index was computed for the original embeddings and for the PCA and UMAP projections under both multiclass and binary label configurations.

III. RESULTS

This section presents the organization of the speaker-level embeddings obtained from the fine-tuned wav2vec 2.0 model. The analysis focuses on the comparison between PCA and UMAP projections, the relationship between the latent space and the fundamental frequency (F_0), and the quantitative evaluation of cluster separability using the Davies–Bouldin index.

A. Speaker-Level Embedding Projections

Figures 1 and 2 show the two-dimensional projections of the speaker-level embeddings using PCA and UMAP, respectively. Each point represents an individual speaker obtained through median aggregation of utterance-level embeddings.

The projections reveal partially separable structures between classes, although substantial overlap remains in some regions of the latent space. In both PCA and UMAP, the *male* class tends to form more compact regions, while the *female* and *other* classes exhibit higher dispersion and stronger overlap.

The UMAP projection produces more visually compact local structures, suggesting stronger preservation of neighborhood relationships. However, PCA preserves a more globally continuous geometric structure. These differences become more evident in the quantitative cluster analysis presented later in this section.

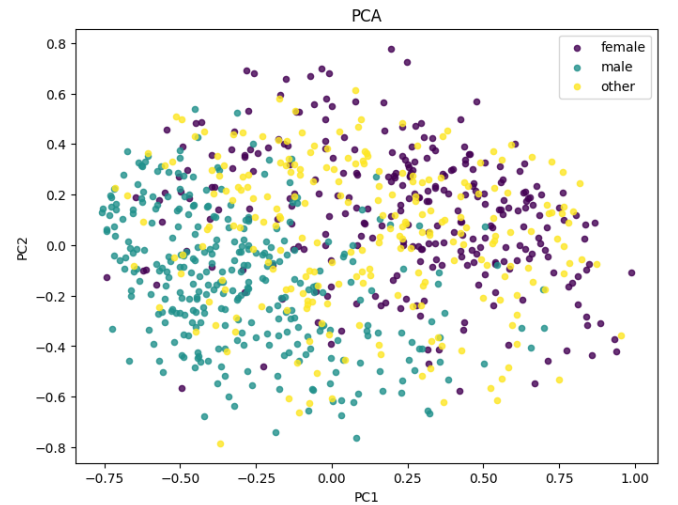


Fig. 1. Speaker-level PCA projection of wav2vec 2.0 embeddings. Colors indicate class labels.

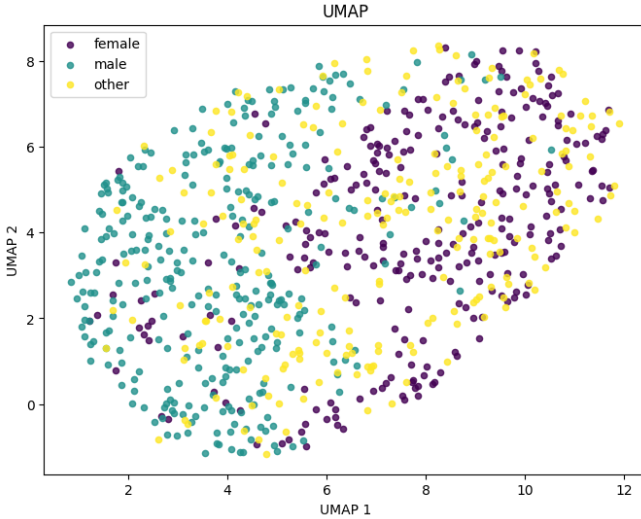


Fig. 2. Speaker-level UMAP projection of wav2vec 2.0 embeddings. Colors indicate class labels.

B. Relationship Between Embeddings and Fundamental Frequency

To investigate whether the learned representations preserve interpretable acoustic information, the PCA and UMAP projections were colored according to the median fundamental frequency (F_0) of each speaker. Figures 3 and 4 show the resulting distributions.

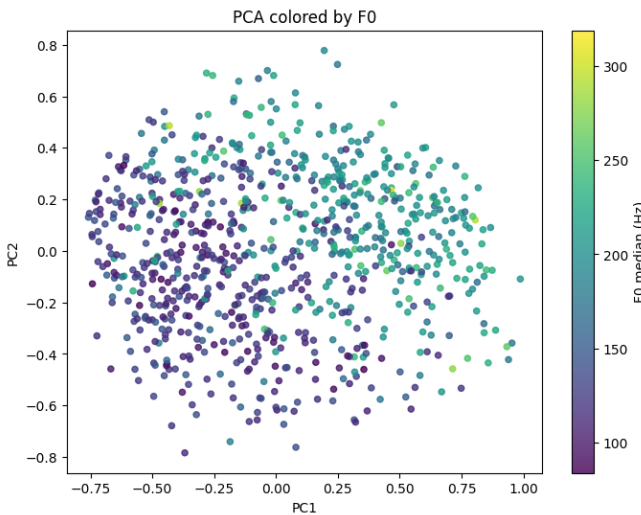


Fig. 3. Speaker-level PCA projection colored according to median F_0 .

Both projections exhibit gradual transitions in F_0 values across the latent space, indicating that the embeddings preserve continuous acoustic variation associated with vocal pitch. Instead of forming discrete regions exclusively associated with specific F_0 ranges, the latent organization appears to follow smooth acoustic gradients.

This behavior suggests that wav2vec 2.0 embeddings encode pitch-related information in a distributed manner. The continuity observed in the projections reinforces the hypothesis that self-supervised speech representations capture acoustic

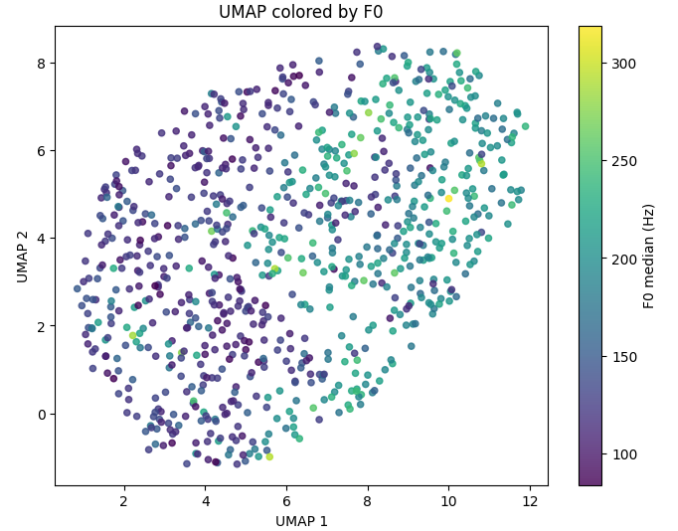


Fig. 4. Speaker-level UMAP projection colored according to median F_0 .

properties beyond explicit class boundaries.

The PCA projection presents smoother global transitions, while UMAP emphasizes local neighborhoods with similar F_0 values. This difference is consistent with the distinct geometric properties of linear and nonlinear dimensionality reduction methods.

C. Davies–Bouldin Analysis

To quantitatively evaluate cluster separability, we computed the Davies–Bouldin (DB) index for the original speaker embeddings and for the two-dimensional PCA and UMAP projections.

Table II summarizes the results for different label configurations.

TABLE II
DAVIES–BOULDIN INDEX FOR DIFFERENT EMBEDDING SPACES AND LABEL CONFIGURATIONS. LOWER VALUES INDICATE BETTER CLUSTER SEPARABILITY.

Configuration	Original	PCA	UMAP
Male / Female / Other	4.10	4.33	5.95
Male vs Female + Other	2.25	1.51	1.60
Male vs Female	1.91	1.26	1.22
Female vs Male + Other	2.79	2.12	2.39

The results indicate that dimensionality reduction improves cluster separability in the binary classification settings, particularly for the *male* versus *female* configuration. In this case, both PCA and UMAP substantially reduce the Davies–Bouldin index relative to the original high-dimensional embedding space.

For the three-class configuration, however, the projections do not improve separability. In particular, UMAP produces a considerably higher DB value, suggesting that the local neighborhood preservation of the method does not necessarily translate into improved global cluster compactness.

Overall, PCA achieved more stable results across different configurations, while UMAP showed stronger sensitivity to the

class structure. These findings suggest that the organization of the speaker-level latent space depends strongly on both the projection method and the label configuration adopted for the analysis.

IV. CONCLUSIONS

This work presented a comparative analysis of linear and nonlinear dimensionality reduction techniques applied to speaker-level speech embeddings extracted from a fine-tuned wav2vec 2.0 model. By comparing PCA and UMAP projections and relating them to the median fundamental frequency (F_0), we investigated how the geometric organization of the latent space reflects acoustic variation across speakers.

The results show that both PCA and UMAP preserve meaningful acoustic organization in the projected embedding spaces. In particular, gradual transitions of F_0 values emerged across the projections, indicating that the learned representations encode continuous pitch-related information rather than strictly discrete class boundaries. PCA produced more stable global structures, while UMAP emphasized local neighborhood organization and compact regional structures.

The Davies–Bouldin analysis further demonstrated that dimensionality reduction improves cluster separability in binary settings, especially in the *male* versus *female* configuration. However, the three-class scenario presented substantially higher overlap, particularly involving the *other* category.

An important observation is that the *other* class consistently appeared geometrically closer to the *female* class than to the *male* class in both the visual projections and the cluster separability analysis. This behavior suggests that the latent organization learned by the model may reflect asymmetric acoustic proximities between classes, raising new hypotheses about how acoustically heterogeneous vocal categories are represented in self-supervised speech embeddings.

More broadly, the overlap between classes and the gradual acoustic transitions observed across the latent space suggest that gender-related vocal information may be represented by the model as a continuous acoustic manifold rather than as strictly discrete categories. In this context, the strong overlap associated with the *other* category may indicate the presence of internally heterogeneous vocal characteristics distributed across different acoustic regions of the embedding space.

Overall, the results reinforce the idea that self-supervised speech embeddings encode multiple interacting acoustic and speaker-related dimensions in a distributed manner. More broadly, this study contributes to the interpretability of deep speech representations through the combined analysis of dimensionality reduction, acoustic information, and latent space geometry.

Future investigations may explore whether these latent geometric structures remain stable across different languages, demographic distributions, and socially heterogeneous vocal datasets.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

REFERENCES

- [1] A. Baeovski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Not specified*, 2020, preprint. [Online]. Available: <https://arxiv.org/pdf/2006.11477.pdf>
- [2] A. Schlichter, “Do voices matter? vocalicity, materiality, gender performativity,” *Body & Society*, vol. 17, no. 1, pp. 31–52, 2011.
- [3] D. H. d. C. Martinho, L. W. Lopes, R. Dornelas, and A. C. Constantini, “Can acoustic measurements predict gender perception in the voice?” *PLOS ONE*, vol. 19, no. 11, pp. 1–22, 11 2024. [Online]. Available: <https://doi.org/10.1371/journal.pone.0310794>
- [4] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer, 2002.
- [5] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” 2020. [Online]. Available: <https://arxiv.org/abs/1802.03426>
- [6] T. Ashihara, M. Delcroix, T. Moriya, K. Matsuura, T. Asami, and Y. Ijima, “What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.17632>
- [7] R. A. Ardila, M. Branson, K. Davis, S. Henretty, J. Kohler, S. Moeller, A. Mulligan, J. Stadermann, C. Townsend, and Y. Weinstein, “Common voice: A massively-multilingual speech corpus,” in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 3513–3517.
- [8] Mozilla Foundation, “Common voice dataset,” <https://commonvoice.mozilla.org>, 2023, accessed: 2025-05-07.
- [9] M. Behlau, *Voz: O Livro do Especialista*, 1st ed. Rio de Janeiro: Revinter, 2013.
- [10] S. M. Paes, N. Ziegler, A.-S. Jotz, A. Tsuji, and L. de Oliveira Cielo, “How does our voice change as we age? a systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age,” *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 2, pp. 533–551, 2020.
- [11] Facebook AI, “facebook/wav2vec2-large-960h-lv60-self,” <https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self>, 2020, hugging Face model card.
- [12] A. Anfossi and D. Chicco, “An easy guide to the davies–bouldin index for unsupervised internal clustering evaluation,” *Discover Computing*, vol. 29, p. 212, 2026. [Online]. Available: <https://doi.org/10.1007/s10791-026-10094-0>