DIME: A Continuous Dialect Distance Measure from Classification Embeddings

Lea Fischbach[1], Alfred Lameli[1], Lucie Flek[2,3]

[1]Research Center Deutscher Sprachatlas, Marburg University, Germany
[2]Bonn-Aachen International Center for Information Technology, University of Bonn, Germany
[3]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
Correspondence: lea.fischbach@uni-marburg.de

Abstract

Dialectal variation in speech is inherently continuous, yet most computational approaches model it as a discrete classification problem. We propose a Dialect Distance Measure derived from embeddings of a dialect classification model (DIME), representing dialectal variation as deviation from a standard reference in embedding space. The score combines distance to a standard reference with projection onto a learned dialect–standard direction. Evaluated on German speech data, the proposed measure shows strong correlation with an independent phonetic reference and remains robust across different reference corpora. In contrast, using the same scoring procedure with off-the-shelf speech embeddings yields less consistent results and does not provide a stable continuous dialectal scale. The suitability of the method for capturing variation is demonstrated using within-speaker as well as between-dialect differences.

1 Introduction

Dialectal variation in speech is inherently continuous, yet most computational approaches treat it as a discrete classification problem, limiting the capture of gradual variation within and across speakers.

In dialectometry, discrete measures such as the Levenshtein distance (Heeringa and Gooskens, 2003; Kessler, 1995; Wieling et al., 2014) quantify phonetic distance to a pre-defined reference. Similar holds for the HS measure (Herrgen et al., 2001), which measures the more fine-grained, i.e. phonetically weighted, distance to Standard German (so-called D-value). Recent work has systematically compared phonetically weighted and unweighted distance measures, showing that explicit phonetic weighting improves dispersion, stability, and interpretability without altering the overall structure of dialect relations (Lameli, 2026). While linguistically grounded, these approaches rely on manual phonetic transcriptions, making them costly, difficult to scale, and sensitive to annotation variability (Cucchiarini, 1996; Bucholtz, 2007; Heeringa et al., 2009; Lameli et al., 2020). Moreover, they primarily capture segmental differences and overlook finer acoustic variation (Bartelds et al., 2020).

Acoustic approaches aim to address these limitations by operating directly on speech signals. Early work by Heeringa et al. (2009) uses engineered features such as formants, while more recent methods employ acoustic distance measures (Bartelds et al., 2020) or articulatory representations (Huang et al., 2025). However, these approaches still rely on feature engineering, alignment procedures, or phoneme-based references.

Machine learning approaches have further explored continuous modeling of dialectal variation using acoustic embeddings and ASR-based features (Johnson et al., 2022; Ghorbani and Hansen, 2024). While effective for prediction tasks, they often depend on intermediate representations such as ASR outputs, introducing additional processing steps and dependencies beyond the acoustic signal.

This raises the question of whether neural speech embeddings themselves encode dialectal variation in a way that allows for a continuous and structured representation.

In this work, we propose an embedding-based dialect distance measure for modeling dialectal variation in representation space. To the best of our knowledge, this is the first measure of distance in the field of dialectology that is derived directly from acoustic signals. The approach builds on speech embeddings from a dialect classification model and interprets dialect intensity as a continuous deviation from a standard reference, combining (i) distance and (ii) projection onto a learned dialect–standard direction. The resulting score can be computed directly from embeddings without requiring phonetic transcriptions or alignment. We evaluate the proposed measure against existing D-

values from the REDE project (Schmidt et al., 2020–), analyze its robustness across reference corpora, compare it to DIME scores derived from off-the-shelf speech representations, and demonstrate its applicability to within-speaker and between-dialect variation.

2 Data

2.1 Speech Data

We conduct our analysis on speech data from the REDE corpus (Schmidt et al., 2020–), a resource for regional German speech across speaking situations, generations, and regions (Fischer and Lameli, 2026). The corpus includes multiple speaking situations designed to capture varying degrees of dialectal and standard language production (Lameli, 2025). In this study, we focus on three conditions that differ systematically in their level of dialectal intention.

The first condition consists of *Wenker sentences in dialect (WSD)*. In this setting, speakers are presented with the 40 "Wenker Sentences" (Wenker, 2013) in Standard German and are asked to translate them into their local dialect. This condition is explicitly *dialect-intended*.

The second condition comprises *Wenker sentences in Standard German (WSS)*. Here, speakers perceive dialectal versions of the Wenker Sentences in their own dialect and are instructed to translate them into their most formal Standard German. This condition is *standard-intended*.

The third condition is the *Conversation among friends (FG)*, in which speakers engage in spontaneous dialogue with a familiar person without the presence of an interviewer. This setting captures natural, uninstructed speech and is assumed to reflect more authentic language use.

Further details on the recording conditions are available in the REDE *Infothek* (Lipfert, 2025). Table 1 summarizes the number of speakers and the total duration of audio data available for each condition, as well as the overall dataset.

Condition	# Speakers	Duration (s)
WSS	540	83,720
WSD	533	83,470
FG	490	85,700
All	558	252,890

Table 1: Overview of the REDE speech data used in this study. The table reports the number of speakers and total duration for each speaking condition: Wenker Sentences in Standard German (WSS), Wenker Sentences in dialect (WSD), and spontaneous conversations among friends (FG).

Corpus	# Speakers	Duration (s)
NWS	640	31,020
Tagesschau (TS)	3	24,870
HUI	15	21,190
Bundestag (BT)	–	20,000

Table 2: Overview of external Standard German corpora used to construct the reference representation.

2.2 External Standard Corpora

To construct a dialect distance measure that quantifies deviation from Standard German, we require a reference representation of standard language. For this purpose, we incorporate multiple external corpora. Table 2 summarizes the number of speakers and the total duration for each external standard corpus.

First, we include recordings of the *The Northwind and the Sun (NWS)* condition from the REDE corpus, in which speakers read a fixed text, providing a controlled approximation of standard pronunciation per speaker. The REDE corpus used in this study contains only male speakers. Consequently, the NWS reference data derived from REDE also consist exclusively of male speakers.

Second, we use a *Tagesschau (TS)* corpus, consisting of speech segments extracted from news presenters of the German television news program. These recordings are expected to represent a highly standardized and formal variety of spoken German. The corpus comprises recordings from three male speakers, with 30–42 *Tagesschau* broadcasts per speaker, covering different recording days between early 2018 and the first quarter of 2020. The corpus is not publicly accessible.

Third, we incorporate data from the *HUI Audio Corpus* (Puchtler et al., 2021), which contains audiobook recordings. We use a subset of the "others" partition and randomly sample male speakers up to a total duration of approximately 20,000 seconds to ensure comparability with other corpora. The restriction to male speakers was chosen to match the REDE data.

Finally, we include speech data from the *German Bundestag (BT) corpus* (Wirth and Peinl, 2023), which consists of recordings from plenary sessions and committee meetings of the German parliament.

We use the "Dirty Delta" subset and randomly sample 12–14-second segments up to 2,000 segments and thus 20,000 seconds. The BT corpus contains speech from both male and female speakers. However, speaker identities and corresponding gender metadata are not consistently available for this dataset.

2.3 Dialectality Reference Measure

As an external reference for dialectal variation, we use the *Dialectality Value (D-value)*, a theoretically motivated measure of phonetically weighted distance originally proposed by Herrgen and Schmidt (1989) and later operationalized in detail by Herrgen et al. (2001). The D-value quantifies the deviation of a regional speech sample from codified Standard German by comparing phonetic realizations with a standardized reference form as described by Lameli (2025).

The method is based on manually annotated IPA transcriptions aligned with codified Standard German reference forms. Phonetic differences are automatically identified and subsequently validated by trained experts, while phenomena not considered indicative of dialectal variation—such as assimilation, reduction, or deletion arising from general speech production—are excluded. The resulting D-value represents the average phonetically weighted deviation from Standard German, averaged across words.

Importantly, the D-value focuses exclusively on segmental phonetic variation (including length). Suprasegmental features such as intonation and prosody are not considered, as they cannot be reliably quantified within the framework. Likewise, syntactic and lexical variation are excluded, making the measure a mainly phonetic representation of dialectality (including—from a morphological perspective—differences in inflection).

The D-value is described in detail in the REDE infothek (Lipfert, 2023). In the context of this study, it serves as an external, linguistically grounded benchmark for evaluating embedding-based dialect intensity measures.

D-values are available for a subset of the REDE recordings used in this study. In total, 525 recordings are annotated with D-values. Of these, 429 could be included in the correlation analyses because a corresponding DIME score was available, including 147 instances of WSS, 136 of WSD, and 146 of FG from 149 speakers.

While the D-value captures phonetically weighted distance based on expert-defined criteria, our approach aims to derive a comparable notion of dialect intensity directly from neural speech embeddings. Crucially, our method is not limited to segmental phonetics and may capture additional sources of variation present in the acoustic signal. Comparing the two measures therefore reveals how strongly embedding spaces encode linguistically meaningful dialectal variation.

3 Method

3.1 Embedding Space

Speech representations are obtained using a two-stage embedding pipeline described in Fischbach et al. (2026) and publicly available on GitHub[1]. In addition, the D-values used in this study, as well as the scripts for the majority of the analysis, are made available at the same repository.

First, all audio recordings are resampled to 16 kHz, converted to mono, and represented using 16-bit resolution and then segmented into non-overlapping chunks of 10 seconds. In the first stage, segment-level embeddings are extracted using *TRILLsson4* (Shor and Venugopalan, 2022a,b), a pre-trained speech representation model developed by Google. This model produces general-purpose acoustic embeddings. In the second stage, the TRILLsson embeddings are passed through a lightweight multilayer perceptron (MLP) that has been trained to classify German dialects. The MLP was trained on speech data derived from the same underlying recordings as the WSD, but using a different segmentation scheme. Rather than using the final classification output, we extract the activations of the last hidden layer as our final representation. These embeddings are 512-dimensional and serve as the basis for all subsequent analyses. This choice allows us to analyze whether representations learned for discrete dialect classification also encode continuous variation.

3.2 Alternative Speech Representations

To assess whether the performance of DIME depends on representations learned specifically for dialect classification, we additionally apply the complete DIME procedure to two off-the-shelf speech representations: Wav2Vec2 (Baevski et al., 2020) and TRILLsson4 (Shor and Venugopalan, 2022a,b).

[1] https://github.com/WoLFi22/DialectClassificationPipeline

TRILLsson4 is also used as the initial feature extractor in the dialect-classification pipeline described in subsection 3.1. For this comparison, DIME is applied directly to the TRILLsson4 representations before they are passed through the dialect-classification MLP. This allows us to assess whether the supervised dialect-classification stage adds information relevant to continuous dialectality estimation.

For Wav2Vec2, we use the `facebook/wav2vec2-base-960h` checkpoint. We extract the contextual representations from the final encoder layer and apply mean pooling across the temporal dimension, resulting in one 768-dimensional representation per 10-second segment.

3.3 Distance-based Component

To quantify dialectal variation, we first define a distance-based measure relative to a reference representation of Standard German.

Let $x \in \mathbb{R}^d$ denote a speech embedding and let c_{std} denote the standard reference centroid, computed as the mean of the segment-level embeddings from the external Standard corpus used in the respective experiment. (When multiple reference corpora are used, their embeddings are pooled before computing the centroid.) We define the dialect distance of x in terms of its distance to this reference:

$$z_{\text{dist}}(x) = \frac{\|x - c_{\text{std}}\|_1 - \mu_{\text{std}}}{\sigma_{\text{std}}} \tag{1}$$

where μ_{std} is the mean L1 distance of the reference embeddings to c_{std}, and σ_{std} is the corresponding standard deviation. Manhattan (L1) distance is motivated by prior work on high-dimensional spaces, which shows that distance measures tend to concentrate as dimensionality increases, reducing distance contrast between samples. Under these conditions, L1 distance has been shown to preserve relative differences more effectively than Euclidean (L2) distance (Aggarwal et al., 2001).

3.4 Directional Component

While the distance-based measure captures only distance from the standard reference, it ignores directional structure in the embedding space. To account for this, we introduce a directional component based on a learned dialect–standard axis.

Let $v \in \mathbb{R}^d$ denote a unit vector representing the direction of maximal dialect–standard separation. This vector is obtained by training a logistic regression classifier on dialect and standard speech embeddings. The resulting weight vector is then normalized to unit length.

We define the projection of an embedding x onto this direction, relative to the standard reference point c_{std}, as:

$$s_{\text{proj}}(x) = \langle x - c_{\text{std}}, v \rangle \tag{2}$$

To obtain a normalized score, we again standardize this value using the distribution of projections of the reference (standard) embeddings:

$$z_{\text{proj}}(x) = \frac{s_{\text{proj}}(x) - \mu_{\text{proj}}}{\sigma_{\text{proj}}} \tag{3}$$

where μ_{proj} and σ_{proj} denote the mean and standard deviation of projections of the standard embeddings onto v. Note that the standard embeddings used to learn v need not be the same as the standard reference embeddings used to compute c_{std} and the projection normalization statistics.

The resulting score $z_{\text{proj}}(x)$ captures dialectal variation along a specific direction in embedding space. By pooling all dialect regions into a single class, we assume that they share a common component of deviation from standard-intended speech. Positive values indicate alignment with the dialectal direction. This formulation allows us to capture dialectal variation not only in terms of magnitude (distance), but also in terms of direction in embedding space.

3.5 Embedding-based Dialect Distance Measure

We combine the distance-based and directional components into a unified dialect distance measure, which we refer to as the *Dialect Distance Measure from Classification Embeddings (DIME)*.

For a given embedding x, the score is defined as:

$$\text{DIME}(x) = w \cdot z_{\text{dist}}(x) + (1 - w) \cdot z_{\text{proj}}(x) \tag{4}$$

where $z_{\text{dist}}(x)$ captures the radial distance from the standard reference, and $z_{\text{proj}}(x)$ captures variation along the dialect–standard direction. The parameter $w \in [0, 1]$ controls the relative contribution of both components. The weighting parameter w allows us to systematically analyze the relative importance of distance-based and directional information. In particular, $w = 1$ corresponds to a purely distance-based measure, while $w = 0$ relies exclusively on the directional component. An overview of the

complete pipeline is provided in Appendix A in Figure 4.

By construction, a value of $\mathrm{DIME}(x) \approx 0$ corresponds to the average realization of the external Standard German reference corpus. Positive values indicate increasing deviation from this reference, while negative values correspond to speech that is more standard-like than the average of the reference corpus. Note that this reference is corpus-dependent and does not represent an absolute notion of standardness.

3.6 Correlation Analysis

To evaluate the relationship between DIME and the reference measure (D-value), we compute correlation analyses between aggregated DIME scores and corresponding D-values.

Since DIME is computed at the segment level, multiple scores are available for each speaker in each recording condition. The number of segments varies across conditions (WSS: 15.74 ± 5.03, WSD: 15.96 ± 4.8, FG: 17.63 ± 13.08). To obtain a robust estimate per speaker and condition, we aggregate segment-level scores using the median. This choice reduces the influence of outliers, such as atypical segments or local acoustic artifacts. The aggregated score is computed for each speaker–condition pair and compared to the corresponding D-value, for which at most one value is available per speaker and condition. Correlation is measured using both Pearson's r and Spearman's ρ.

To analyze the relative contribution of the distance-based and directional components, we vary the weighting parameter w over 21 values ($w \in [0, 1]$, step size 0.05). For each value, correlation with the D-value is computed, and the best-performing weight is used for subsequent analyses.

Particular attention is given to the FG condition. The projection direction v is learned using a logistic regression model trained on WSD and either WSS or the pooled external standard corpora, depending on the setting. As FG is not used to learn this direction, it provides an evaluation scenario independent of the projection model and therefore provides a stronger test of generalization.

4 Evaluation

4.1 Effect of Direction vs. Distance

We first analyze the relative contribution of the distance-based and directional components in Equation 4. For both settings, the standard reference centroid and the corresponding normalization statistics are computed from the pooled embeddings of all external standard corpora (HUI, NWS, TS, and BT). We vary only the data used to learn the projection direction v to investigate how the choice of data used to learn v affects the resulting correlation with the D-value.

To this end, we compare two settings for training the logistic regression model used to estimate v: (i) using standard-intended recordings from the same corpus as the dialect data (WSS), and (ii) using all external standard corpora (HUI, NWS, TS, BT). For each setting, we vary the weighting parameter w and determine the value that maximizes the correlation with the D-value.

When the projection direction is learned from the pooled external standard corpora and WSD, the optimal weight is $w = 0.95$, yielding a Pearson correlation of $r = 0.53$, $p < .001$. Performance is therefore driven almost entirely by the distance-based component. In contrast, when the projection direction is learned from WSS and WSD, the optimal weight shifts to $w = 0.35$, and the Pearson correlation increases substantially to $r = 0.66$, $p < .001$. In this setting, the directional component therefore contributes substantially to the combined score.

As shown in Figure 1, Pearson and Spearman correlations exhibit the same pattern. The WSS–WSD direction produces a clear maximum at intermediate weights ($w \approx 0.35$), confirming that neither component alone is optimal and that performance degrades when either dominates, whereas correlations in the external-standard setting increase as more weight is assigned to the distance component.

This pattern is also reflected in the performance of the individual components. For the WSS–WSD direction, the distance-based score alone yields a correlation of $r = 0.53$, whereas the directional score alone reaches $r = 0.63$. Although the directional component thus captures more of the relevant structure than the distance-based component, combining both produces the highest correlation.

To further assess whether these correlations could arise by chance, we compare the distance-based and directional components against separate null models. For the distance component, the pooled external standard center yields a Pearson correlation of $r = .531$ with the D-value. Random centers sampled from the pooled embedding space or from a Gaussian approximation of that space

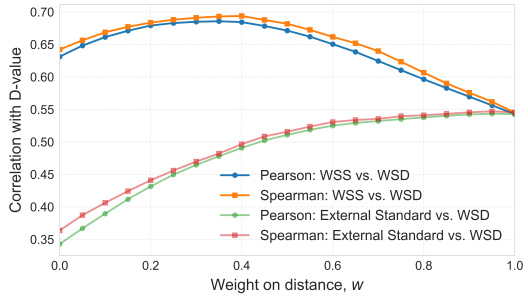

Figure 1: Correlation with the D-value as a function of the weighting parameter w for projection directions learned from WSS versus WSD and from the pooled external standard corpora versus WSD. In both settings, the standard centroid and normalization statistics are computed from the pooled external standard corpora (HUI, NWS, TS, and BT).

yield lower correlations on average ($M_r = .323$ and $M_r = .317$, respectively). However, approximately 10–11% of the random centers achieve correlations at least as high as the external standard center. Thus, the selected center performs better than random centers on average, although its performance is not unique within the embedding space. For the directional component, projection onto the learned WSS–WSD direction yields a Pearson correlation of $r = .631$. The distribution obtained from 1,000 random unit directions is centered near zero ($M_r = .002$, $SD = .238$, 95% interval $[-.421, .428]$). Although individual random directions can produce moderate positive or negative correlations, none reaches the absolute correlation of the learned direction (empirical two-sided $p = .001$). This indicates that the learned WSS–WSD direction captures a systematic dialectality-related gradient that is not obtained along arbitrary directions in the embedding space.

4.2 Robustness

Table 3 summarizes the correlation between DIME and the D-value for different choices of the standard reference corpus. Results are reported overall and separately for each speaking condition (WSS, WSD, FG). In addition, we report the AUC for distinguishing WSD and WSS segments (separation) as well as the optimal weighting parameter w.

Overall correlations are stable across reference corpora (r = 0.64–0.69). The BT corpus yields the highest correlation in the FG condition ($r = 0.583$) and also performs best overall ($r = 0.686$). This improvement is primarily driven by a substantially higher correlation in the WSS condition ($r = 0.474$), which appears to be the most challenging setting, where the difference to other corpora is most pronounced.

Despite these differences, the results indicate a high degree of robustness with respect to the choice of the reference corpus. Correlations, separation performance, and optimal weights vary only moderately across all settings. Interestingly, the separation performance (AUC) remains remarkably stable across all reference corpora, indicating that the ability to distinguish dialectal from standard speech is largely independent of the chosen reference.

The optimal weight w consistently lies in the range of 0.15 to 0.35, indicating that the directional component is generally more informative than the distance-based component across all reference choices. Combining multiple standard corpora (e.g., TS + BT or all corpora) does not lead to systematic improvements. Instead, the resulting correlations tend to fall between those of the individual corpora. This suggests that increasing the amount and diversity of standard reference data does not necessarily enhance the alignment with the D-value.

In addition to varying the reference corpus, we investigate whether the performance of DIME depends on representations learned specifically for dialect classification. We therefore apply the same DIME scoring procedure to off-the-shelf Wav2Vec2 (Baevski et al., 2020) and TRILLsson4 representations, using the BT corpus as the external standard reference. As shown in Table 3, DIME applied to these alternative representations yields moderate to strong overall correlations with the D-value, particularly for dialect-intended speech (WSD). However, the condition-specific results are less consistent: correlations for WSS are close to zero or negative, whereas correlations for WSD are substantially higher. This suggests that off-the-shelf speech representations can support the separation of speaking conditions, but not a stable notion of dialect distance to a standard reference.

As an additional out-of-domain validation, we apply the score to TS recordings using the BT corpus as standard reference and a projection direction learned from WSD and WSS. Aggregating scores at the recording level yields predominantly negative values (median = -0.15, IQR [-0.24, -0.03], n=103), indicating that TS recordings receive consistently lower DIME scores than the BT reference distribution. This behavior is in line with expectations, as TS speech is generally considered more

	Wav2Vec2	TRILL	TS	HUI	BT	NWS	TS+BT	All
Overall	.527***	.685***	.657***	.643***	.686***	.652***	.669***	.660***
WSS	-0.079	.069	.390***	.343***	.474***	.346***	.428***	.413***
WSD	.412***	.518***	.542***	.514***	.570***	.540***	.552***	.546***
FG	.149	.402***	.515***	.478***	.583***	.487***	.543***	.527***
AUC	.820	.969	.804	.809	.793	.806	.801	.797
Best w	0.00	0.00	0.20	0.15	0.35	0.25	0.30	0.35

Table 3: Pearson correlation between DIME and the D-value for different choices of the standard reference corpus and underlying representations. Results are reported overall and per condition (WSS, WSD, FG). For the alternative speech representations (Wav2Vec2 and TRILLsson4), the Bundestag corpus (BT) is used as the standard reference. AUC denotes the area under the curve for distinguishing WSD vs. WSS segments. The optimal weighting parameter w is reported for each setting. *Note:* * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

standardized—and somewhat artificial—than parliamentary speech, suggesting that the proposed score generalizes to unseen data distributions.

4.3 Within-Speaker Variation

To further analyze how the proposed score behaves within speakers across different speaking conditions, we perform a within-speaker analysis and compare the DIME to the D-value. This type of analysis corresponds to what is often referred to as *vertical variation* in dialectology (Schmidt, 2010).

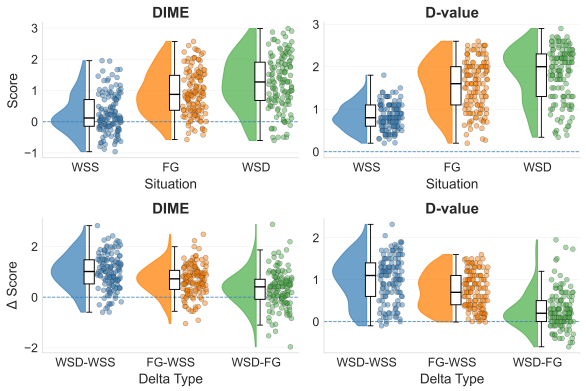

Figure 2: Within-speaker variation of the DIME (left) and the D-value (right), computed on the subset of speakers for which both measures are available in all three situations. Top: Distribution of scores per speaking condition (WSS, FG, WSD). Bottom: Pairwise within-speaker differences between conditions.

Figure 2 shows the distributions of both measures. The upper row presents the absolute scores per condition (WSS, FG, WSD) for each speaker, while the lower row shows pairwise differences between conditions (e.g., WSD − WSS) computed per speaker. These differences (deltas) allow for a direct comparison of how individual speakers shift between speaking situations, independent of the overall scale of the measures. As the BT corpus yielded the strongest overall correlations, it is used here as the standard reference for computing DIME. Since D-values are not available for all speakers and speaking situations, both measures are restricted to the subset of 133 speakers for whom both measures are available in all three speaking situations, yielding a balanced set of 399 observations per measure across WSS, FG, and WSD.

Across both measures, a consistent ordering is observed (WSD > FG > WSS), holding for both absolute scores and within-speaker differences. However, the scales of the two measures differ fundamentally. The D-value has an interpretable absolute zero corresponding to near-standard pronunciation, whereas DIME is standardized relative to the reference distribution. Negative DIME values therefore indicate a below-average combination of distance and directional projection relative to the standard reference corpus.

Delta distributions provide a more direct view of within-speaker variation across conditions. For both measures, the pairwise differences WSD − WSS and FG − WSS are predominantly positive, indicating more dialectal speech in dialect-intended and spontaneous settings. In contrast, the WSD − FG difference is centered closer to zero, suggesting only minor differences between dialect-intended and spontaneous speech. However, the comparatively large spread of the distributions indicates substantial variability across speakers, with some speakers producing spontaneous speech that is close to dialectal production, while others show larger deviations.

While both measures exhibit similar qualitative patterns, the D-value shows a relatively broader spread for the WSD − WSS difference, indicating greater variability across speakers in the magnitude of the shift between standard-intended and

dialect-intended speech. In comparison, DIME distributions appear more compact, suggesting more homogeneous within-speaker variation, although differences in scale between the measures limit direct comparability.

4.4 Between-Dialect Structure

As speakers differ substantially in their reported ability to speak the dialect of their home region, we restrict the analysis to speakers whose self-assessed dialect proficiency lies above the global mean rating. Previous work has shown that very low self-assessed dialect proficiency can negatively affect dialect classification performance, whereas this effect was no longer observed above the mean rating threshold (Fischbach et al., 2026). The filtering step is therefore intended to reduce the influence of speakers whose speech exhibits only low levels of dialectality. We consider the dialect regions that were also used to train the MLP described in subsection 3.1. Dialect regions represented by fewer than five speakers after filtering are excluded from the analysis. For each speaker, we first compute the median DIME across segments within a given speaking condition. Subsequently, these speaker-level scores are aggregated by taking the median across all speakers belonging to the same dialect region. This results in one representative score per dialect and condition, allowing for a direct comparison of dialect regions in terms of their distance to the chosen standard reference.

Figure 3 shows the resulting dialect-level scores for the dialect-intended condition (WSD), using the BT corpus as the standard reference in orange. The results indicate that High Alemannic exhibits the highest median distance from the reference, while Brandenburgian shows the lowest. This ordering broadly aligns with previous work on regional dialectality in German, which reports comparatively low dialect intensity in eastern Central German regions such as Brandenburg and stronger dialect intensity in many northern and southern dialect regions (Lameli, 2025, p. 63).

To assess the influence of the chosen reference corpus on the resulting dialect scores, we repeat the analysis using the TS corpus as reference. As TS speakers operate within a highly standardized broadcast domain and are expected to adhere closely to codified pronunciation norms (Krech et al., 2009), their speech is commonly used as an empirical approximation of Standard German (Torres and Müller, 2021). The corpus can therefore be

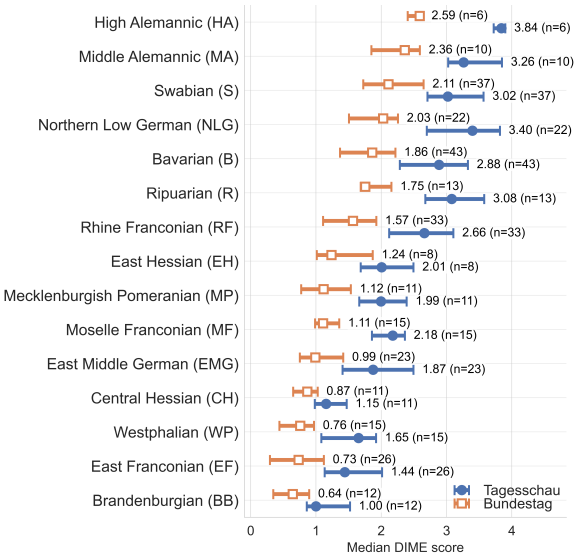

Figure 3: Median DIME scores by dialect region in the WSD condition for two standard reference corpora. Scores are first aggregated per speaker and subsequently summarized per dialect. Markers and adjacent values denote the median, while horizontal intervals show the interquartile range (Q_1–Q_3) across speakers. Numbers in parentheses indicate the number of unique speakers.

considered a stronger proxy for Standard German.

Using this reference shifts dialect scores towards higher values, indicating greater distance from the chosen standard reference. This is also shown in Figure 3 (blue). However, this shift is not uniform across dialects. While the overall ordering remains broadly similar, some relative differences change. For instance, Northern Low German is ranked as more distant from the standard than Swabian with the TS reference, whereas the opposite pattern holds with the BT reference. This highlights that the choice of reference corpus affects not only the absolute scale but also the relative positioning of dialects.

Although the BT corpus yields higher correlations with the D-value, TS provides a more linguistically plausible reference for standardness, indicating a trade-off between empirical agreement with the D-value and the linguistic plausibility of the reference corpus.

5 Discussion

Dialectal variation is primarily organized along a directional gradient in embedding space. While distance captures global deviation from the standard reference, the projection component provides the more informative signal, as reflected by consistently low optimal weights ($w \in [0.15, 0.35]$).

The best performance is achieved by combining both components. However, the projection is only effective when learned from domain-matched data. Because WSS and WSD involve the same speakers and the same set of Wenker-sentences, their primary systematic difference is intended dialectality; this controlled contrast likely yields a more consistent separation signal than external standard corpora, although effects associated with the direction of the translation task cannot be fully excluded.

Across reference corpora, the proposed score remains robust: correlations with the D-value are consistently high, while separation performance (AUC) is largely unaffected by the choice of reference. The comparatively high AUC values are partly expected, as the directional component is learned by logistic regression to separate WSS from WSD and receives substantial weight in the final score. AUC therefore primarily reflects the linear separability of the two conditions along the learned direction. Correlation with the D-value imposes a stricter requirement, as the scores must additionally preserve graded differences in dialectality between speakers. This helps explain why alternative representations can distinguish WSS and WSD without yielding an equally meaningful continuous dialectality scale. The optimal weighting parameter is also stable, indicating that the balance between distance and direction is not highly sensitive to the specific corpus. Combining multiple reference corpora does not yield systematic improvements, suggesting that increasing the amount of standard data alone does not improve alignment with the D-value.

Although correlations are consistently significant, they remain below perfect agreement, reflecting that both measures capture different aspects of dialectal variation. While the D-value is based on manually annotated segmental differences relative to a codified standard, DIME operates directly on acoustic embeddings and may additionally capture suprasegmental cues such as intonation and prosody, although their contribution is not isolated in the present analysis. Evidence for out-of-domain generalization is also provided by the consistent shift towards more standard-like values for TS data using BT as standard reference. However, this analysis is limited by the absence of an external reference measure.

The within-speaker analysis further confirms that both DIME and D-value capture systematic variation across speaking conditions (WSD > FG > WSS). At the same time, DIME exhibits more compact distributions, whereas the D-value shows greater variability across speakers, likely reflecting differences in the nature and scaling of both measures.

At the dialect level, the scores allow direct comparison of dialect regions in embedding space and suggest a largely continuous structure of dialect variation relative to the chosen standard reference. However, both absolute values and ordering depend on the reference corpus. While BT yields the highest correlation with the D-value, TS provides a more linguistically plausible approximation of Standard German, highlighting a trade-off between empirical alignment and theoretical assumptions about standardness. Differences in sample size and the degree of dialect production should also be considered, as both may affect the stability of dialect-level estimates.

Overall, the findings suggest that neural speech embeddings encode dialectal variation as a graded geometric structure. Embeddings derived from a dialect classification model therefore contain not only discrete class information but also information about relative dialect distance. Unlike approaches based on ASR or handcrafted features, the proposed method operates directly on learned acoustic representations.

6 Conclusion

In this work, we introduced a dialect distance measure from classification embeddings (DIME) that models dialectal variation as a continuous geometric structure in representation space. By combining distance to a standard reference with projection onto a learned dialect–standard direction, the measure captures both global deviation and structured variation in speech.

Our results show that embeddings from a dialect classification model encode meaningful continuous information about dialectal variation. The proposed score correlates well with the D-value as an independent phonetic reference, is robust across reference corpora, and supports the analysis of both within-speaker and between-dialect variation. Future work will investigate dynamic variation within conversations and speaker-specific dialectal profiles.

Limitations

A limitation of the present study is that the embedding model was trained on data closely related to the evaluation setting, which may have influenced the observed structure. Future work should investigate whether similar patterns hold for embeddings trained on more diverse or unrelated data. Separately, DIME relies on representations learned from supervised dialect labels. Although new recordings can subsequently be scored without transcriptions or dialect labels, the resulting geometry may depend on the quality, granularity, and composition of the training labels. Generalizability is further limited by speaker composition: REDE contains only male speakers, and the TS and HUI subsets were matched accordingly. BT includes both male and female speech, but their proportions are unknown. It therefore remains unclear whether DIME behaves comparably for female speakers.

Acknowledgments

This research is supported by the Federal Ministry of Research, Technology and Space (BMFTR) (grant AnDy 16DKWN007) and the Academy of Sciences and Literature Mainz (grant REDE 0404), the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence LAMARR22B, and the Research Center Deutscher Sprachatlas Marburg. We are grateful to three anonymous reviewers for helpful comments.

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Martijn Bartelds, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2020. A new acoustic-based pronunciation distance measure. *Frontiers in Artificial Intelligence*, 3:39.

Mary Bucholtz. 2007. Variation in transcription. *Discourse Studies*, 9(6):784–808.

Catia Cucchiarini. 1996. Assessing transcription agreement: methodological aspects. *Clinical Linguistics & Phonetics*, 10(2):131–155.

Lea Fischbach, Alfred Lameli, and Lucie Flek. 2026. Beyond accuracy: Analyzing dialect confusion in automatic speech-based dialect classification. In *Proceedings of the First Workshop on Dialects in NLP – A Resource Perspective*, pages 93–103, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Hanna Fischer and Alfred Lameli. 2026. German dialects across situations, generations, and regions: The REDE corpus as an oral resource for NLP. In *Proceedings of the First Workshop on Dialects in NLP – A Resource Perspective*, pages 144–152, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Shahram Ghorbani and John HL Hansen. 2024. Advanced accent/dialect identification and accentedness assessment with multi-embedding models and automatic speech recognition. *The Journal of the Acoustical Society of America*, 155(6):3848–3860.

Wilbert Heeringa and Charlotte Gooskens. 2003. Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities*, 37(3):293–315.

Wilbert Heeringa, Keith Johnson, and Charlotte Gooskens. 2009. Measuring Norwegian dialect distances using acoustic features. *Speech Communication*, 51(2):167–183.

Joachim Herrgen, Alfred Lameli, Stefan Rabanus, and Jürgen Schmidt. 2001. Dialektalität als phonetische Distanz. Ein Verfahren zur Messung standarddivergenter Sprechformen.

Joachim Herrgen and Jürgen Erich Schmidt. 1989. Dialektalitätsareale und Dialektabbau. In Wolfgang Putschke, Werner H. Veith, and Peter Wiesinger, editors, *Dialektgeographie und Dialektologie. Günter Bellmann zum 60. Geburtstag von seinen Schülern und Freunden*, volume 90 of *Deutsche Dialektgeographie*, pages 304–346. Elwert, Marburg.

Kevin Huang, Sean Foley, Jihwan Lee, Yoonjeong Lee, Dani Byrd, and Shrikanth Narayanan. 2025. On the relationship between accent strength and articulatory features. In *Interspeech 2025*, pages 4538–4542.

Alexander Johnson, Kevin Everson, Vijay Ravi, Anissa Gladney, Mari Ostendorf, and Abeer Alwan. 2022. Automatic dialect density estimation for African American English. In *Interspeech 2022*, pages 1283–1287.

Brett Kessler. 1995. Computational dialectology in Irish Gaelic. In *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics.

Eva-Maria Krech, Eberhard Stock, Ursula Hirschfeld, and Lutz-Christian Anders. 2009. *Deutsches Aussprachewörterbuch*. De Gruyter Mouton, Berlin, New York.

Alfred Lameli. 2025. Gesprochenes Deutsch in den Regionen: Eine Standortbestimmung für die Bundesrepublik Deutschland. In Nadine Proske, Thilo Weber, Monika Dannerer, and Arnulf Deppermann, editors, *Gesprochenes Deutsch: Struktur, Variation, Interaktion*, pages 51–80. De Gruyter, Berlin, Boston.

Alfred Lameli. 2026. Evaluating phonetically weighted and unweighted distance measures in dialectometry. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4141–4151, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Alfred Lameli, Elvira Glaser, and Philipp Stöckle. 2020. Drawing areal information from a corpus of noisy dialect data. *Journal of Linguistic Geography*, 8(1):31–48.

Salome Lipfert. 2023. REDE-Infothek: Dialektalitätswertmessung. Accessed: 2026-04-16.

Salome Lipfert. 2025. REDE-Infothek: REDE-Aufnahmesituationen. Accessed: 2026-04-16.

Pascal Puchtler, Johannes Wirth, and René Peinl. 2021. HUI-Audio-Corpus-German: A high quality TTS dataset. In *KI 2021: Advances in Artificial Intelligence*, pages 204–216, Cham. Springer International Publishing.

Jürgen Erich Schmidt. 2010. Language and space: The linguistic dynamics approach. In Peter Auer and Jürgen Erich Schmidt, editors, *Language and Space: An International Handbook of Linguistic Variation. Volume 1: Theories and Methods*, volume 30.1 of *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*, pages 201–225. De Gruyter Mouton, Berlin/New York.

Jürgen Erich Schmidt, Joachim Herrgen, Roland Kehrein, Alfred Lameli, and Hanna Fischer. 2020–. Regionalsprache.de: Forschungsplattform zu den modernen Regionalsprachen des Deutschen. Digitale Sprachressource. Bearbeitet von Lisa Dücker, Robert Engsterhold, Marina Frank, Heiko Girnth, Simon Kasper, Juliane Limper, Salome Lipfert, Georg Oberdorfer, Tillmann Pistor, Anna Wolańska. Unter Mitarbeit von Dennis Beitel, Lea Fischbach, Milena Gropp, Heiko Kammers, Maria Luisa Krapp, Vanessa Lang, Salome Lipfert, Nathalie Mederake, Jeffrey Pheiff, Bernd Vielsmeier. Studentische Hilfskräfte.

Joel Shor and Subhashini Venugopalan. 2022a. TRILLsson: Distilled Universal Paralinguistic Speech Representations. In *Interspeech 2022*, pages 356–360.

Joel Shor and Subhashini Venugopalan. 2022b. TRILLsson: Distilled universal paralinguistic speech representations [pre-trained model].

Adriana Rosalina Galván Torres and Tanja Müller. 2021. Die Tagesschau in 100 Sekunden: Die Aussprache von <ä> bei Nachrichtensprecher*innen. *Sincronía*, 25(80):803–828.

Georg Wenker. 2013. *Schriften zum "Sprachatlas des Deutschen Reichs". Gesamtausgabe*. Number 111.1–3 in Deutsche Dialektgeographie. Olms, Hildesheim, New York, Zürich. Unter Mitarbeit von Johanna Heil und Constanze Wellendorf.

Martijn Wieling, Jelke Bloem, Kaitlin Mignella, and Mona Timmermeister. 2014. Measuring foreign accent strength in English: Validating levenshtein distance as a measure. *Language Dynamics and Change*, 4:253–269.

Johannes Wirth and René Peinl. 2023. ASR Bundestag: A large-scale political debate dataset in German. In *Proceedings of SAI Intelligent Systems Conference*, pages 190–202. Springer.

A Overview of the DIME Pipeline

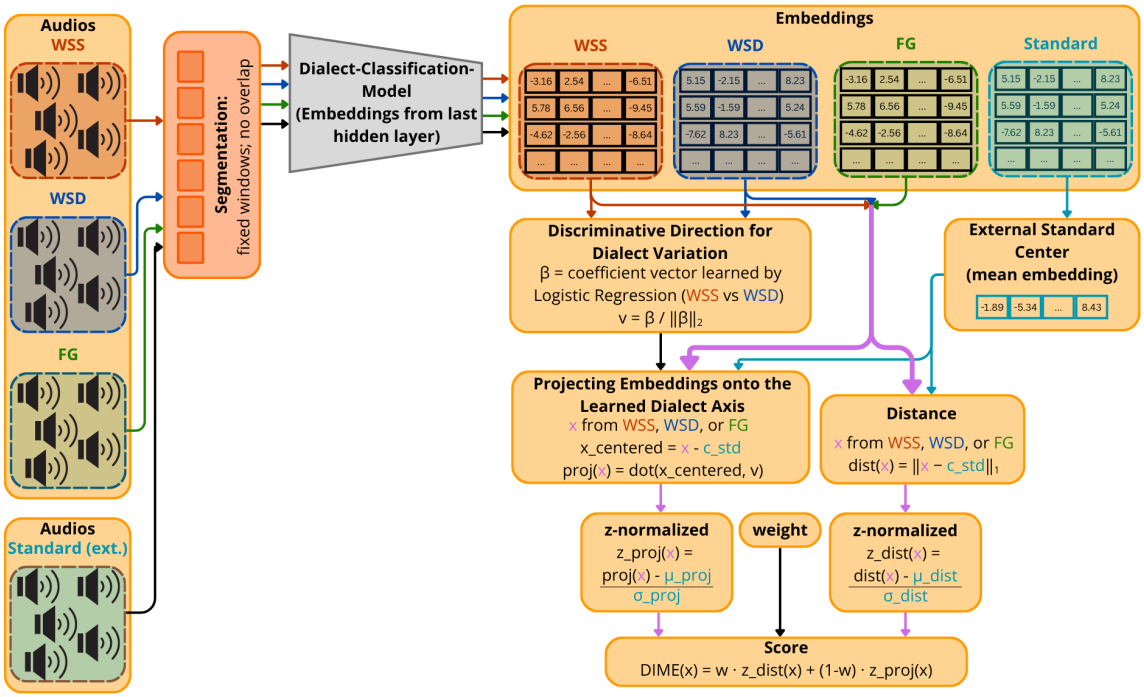

Figure 4: Overview of the DIME pipeline. Segment-level classification embeddings are used to construct an external standard center and to learn a WSS–WSD direction. The z-normalized distance and projection components are subsequently combined into the final DIME score.