Beyond Accuracy: Analyzing Dialect Confusion in Automatic Speech-Based Dialect Classification

Lea Fischbach[1], **Alfred Lameli**[1], **Lucie Flek**[2,3]
[1]Research Center Deutscher Sprachatlas, Marburg University, Germany
[2]Lamarr Institute for ML and AI, Germany
[3]b-it Center, University of Bonn, Germany
lea.fischbach@uni-marburg.de

Abstract

Automatic dialect classification is commonly treated as a supervised task with a primary focus on overall accuracy. In this paper, we argue that classification errors and model uncertainty provide valuable insights into dialectal structure and variation. We analyze a speech-based dialect classification model trained on German dialect data from three generations and evaluated across 250 speaker-disjoint splits (median weighted $F_1 = 0.42$). A systematic confusion analysis shows that misclassifications are largely explained by speaker diversity, dialectal similarity, geographical proximity, and speaker self-assessment. Among these factors, the number of speakers per dialect has the strongest impact on performance, while frequent confusions between closely related dialects reflect inherent linguistic similarity rather than model limitations. Generational analyses further indicate that younger speakers exhibit reduced dialectal distinctiveness, although core dialectal features remain shared across generations. By explicitly modeling classification uncertainty, the proposed approach enables the analysis of dialect transition areas and gradient dialect boundaries. Overall, this work demonstrates that automatic dialect classification can serve not only as a predictive task but also as a tool for dialectological analysis.

Keywords: automatic dialect classification, dialect confusion, dialect continua, generational variation

1. Introduction

Automatic dialect classification from speech is commonly framed as a supervised classification task (Hosseini-Kivanani et al., 2025; Ghafoor et al., 2025). While this perspective emphasizes overall performance, a closer examination of classification errors, confusions, and low-confidence predictions reveals important information about dialectal structure, speaker variation, and the gradient nature of dialect boundaries.

In line with this view, several studies have pointed out structural challenges inherent to dialect classification. Jokisch and Dobbriner (2019), for example, classified nine German dialects using acoustic features such as MFCCs. They report that some dialects and dialect groups are inherently similar and therefore difficult to distinguish, even for human listeners. In addition, they note that some speakers in their test corpus articulate close to Standard German, which may conflict with the forcibly assigned dialect labels used during training. Further challenges include unequal speaker distributions across dialects, unbalanced dialect region sizes, and sparse training data for certain dialect groups. While these factors are clearly identified, their individual impact on misclassification behavior is not examined in detail.

Similarly, Stucki and Randjelovic (2021) investigated the automatic classification of 21 Swiss German dialects using Wav2Vec-XLSR-53 representations. They observed that some dialects were recognized reliably, whereas others were almost entirely misclassified. Moreover, misclassified samples were predominantly assigned to geographically neighboring dialects, indicating dialectal similarity. At the same time, they report that a small number of speakers contributed a large proportion of samples for some dialects, raising the risk that the model learned speaker-specific characteristics rather than dialect features.

In addition, Misganaw and Roller (2022) classified 7,814 short audio snippets from eight dialects in Middle West Germany using a multilabel SVM. However, their dataset comprised only one to two speakers per dialect, with the same speakers used for training and testing. Under these conditions, high performance likely reflects speaker recognition rather than dialect classification, a factor that was not explicitly addressed in their study.

Taken together, these studies demonstrate that reported classification results are strongly influenced by the choice and similarity of dialects and dialect groups, the number and distribution of speakers, and the degree to which speaker-specific variation is controlled (Darjaa et al., 2018). Despite these observations, systematic analyses that disentangle the relative contributions of these factors to misclassification behavior remain rare, particularly with respect to quantitatively separating speaker-related effects from dialectal similarity under speaker-disjoint evaluation settings. While un-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 93–103
16 May 2026.

supervised and data-driven approaches aim to derive similarity structures from acoustic data or predefined linguistic features, for example through distributional similarity metrics such as token distribution similarity (TDS) (Gogoulou et al., 2024) or dialectometric clustering based on linguistic distance measures (Sciarretta, 2024), they do not explicitly model misclassification behavior with respect to predefined dialect categories or allow for a direct analysis of how similarity interacts with speaker variation and data-related factors. In this paper, rather than treating misclassifications solely as model failures, we interpret systematic confusions as potentially informative signals reflecting underlying dialectal relationships, speaker distributions, and sociolinguistic variation. This perspective aligns with previous work on dialect classification, which emphasizes that low performance may arise from dialectal similarity, speaker heterogeneity, and imbalanced data distributions rather than from limitations of the classification model alone (Jokisch and Dobbriner, 2019; Ferragne and Pellegrino, 2007; Stucki and Randjelovic, 2021).

With our study, we therefore do not aim to achieve maximal classification accuracy. Instead, we analyze a speech-based dialect classification model with the goal of understanding systematic misclassifications and low-confidence predictions. We show how such patterns can be related to sociolinguistic factors (generation and speaker self-assessment), geolinguistic properties (dialectal similarity and transition areas), and data-related aspects (speaker number). By doing so, we demonstrate how dialect classification models can be used as analytical tools to investigate dialect variation beyond hard class assignments.

The contributions of this paper are as follows: (i) we present a speech-based dialect classification model trained on a large German dialect corpus; (ii) we provide a systematic analysis of confusions and misclassifications, identifying four main contributing factors; (iii) we introduce an approach to exploit model uncertainty for the analysis of dialect transition areas.

2. Data and Annotation

The audio data used in this study originates from the REDE project (Schmidt et al., 2020–), a large-scale project on regional linguistic variation in Germany. We focus on a subset of the corpus in which speakers were asked to translate 40 Standard German sentences into their local dialect.

The dataset comprises recordings from three generational groups, defined by age: older speakers (60+), middle-aged speakers (42–59), and younger speakers (17–26), with exact age available for 516 speakers (≈90%). The distribution of unique speakers as well as the total duration of audio per generation used in this study are summarized in Table 1. The corresponding age distribution of speakers is shown in Figure 1, illustrating both the separation between the predefined age groups and the variability within them.

Generation	#Speakers	Usable audio (s)
Older speakers	198	63520
Middle-aged speakers	237	58750
Younger speakers	139	31700

Table 1: Number of unique speakers and total duration of usable audio (in seconds) per generational group included in this study.

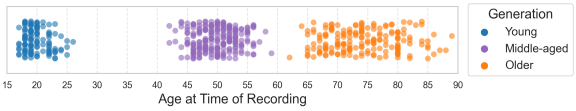

Figure 1: Age distribution of speakers across the three generational groups (older, middle-aged, younger) in the dataset. Each point represents an individual speaker age at the time of recording.

Figure 2 illustrates the geographical distribution of the dialect areas and their transition zones based on Wiesinger (1983), as well as the recording locations covered in the dataset. Dark blue points indicate recordings from core dialects, each associated with a specific dialect label, while light blue points represent recordings from transition zones. These transition zones are of particular interest, as they reflect the gradient nature of dialectal variation rather than strictly discrete boundaries. Dialect areas shown in gray were excluded from the analysis due to an insufficient number of speakers.

2.1. Speaker Self-Assessment

In addition to the speech recordings, speaker self-assessments were collected for each speaker. The self-assessment data were gathered through a standardized questionnaire[1]. Speakers were asked, among others, to rate their active (spoken) competence in the local dialect. The relevant question was formulated as follows: “How well can you speak the [placeholder for the term used by the informants to refer to the way of speaking of long-established residents in the locality] of your hometown?” Responses were provided on a rating scale ranging from 0 (*not at all*) to 6 (*perfect*), with increments of 0.5. The collected data are available in tabular form[2].

[1] available at https://doi.org/10.57712/lingurep-59870
[2] https://doi.org/10.57712/lingurep-59867

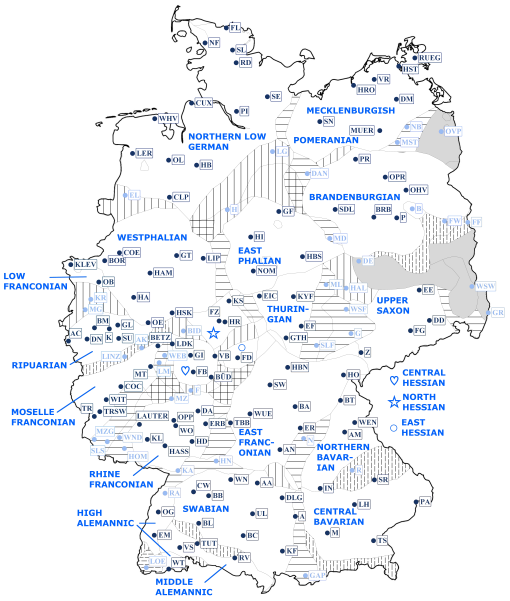

Figure 2: Dialect areas, transition zones according to Wiesinger (1983), and survey locations in the dataset. White areas represent labeled dialect regions, hatched areas transition zones, and gray areas regions excluded from analysis. Points indicate survey locations with location abbreviations.

3. Model Architecture and Training

Figure 3 provides an overview of the important steps from the training pipeline. For model training, only recordings from dialect areas were used. Speech samples originating from transition zones were excluded from the training process and are considered exclusively in the analysis.

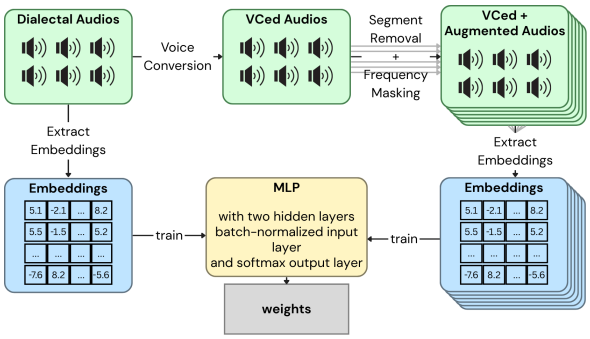

Figure 3: Overview of the training pipeline.

All audio recordings were first subjected to voice conversion, mapping all audios to a single target speaker voice. Voice conversion is not used to improve absolute performance, but to isolate dialect-related variation by minimizing speaker-specific acoustic variability. This step is motivated by the observation that voice conversion reduces speaker-related variability other than dialectal and thereby enables the model to focus more strongly on dialect-specific linguistic and phonetic features (Fischbach et al., 2025a) while preserving linguistic content (Sisman et al., 2020). To further increase data variability and robustness, six augmented versions were generated for each voice-converted audio file using frequency masking and segment removal as described in (Fischbach et al., 2025b).

All resulting audio files were subsequently cut into fixed-length segments of 10 seconds and normalized by resampling to 16 kHz, converting to 16-bit depth, and mixing down to mono. Incomplete final segments shorter than 10 seconds were discarded. Given the total duration per generation reported in Table 1, the number of resulting segments follows directly from this fixed windowing procedure. Speech embeddings were extracted from each segment using Google's TRILLsson 4 model (Shor and Venugopalan, 2022) and used to train a multilayer perceptron (MLP) classifier. The MLP consists of a batch-normalized input layer followed by two fully connected hidden layers with He-normal initialization, L2 kernel and L1 activity regularization, and LeakyReLU activations. Dropout is applied after each hidden layer to reduce overfitting. The output layer uses a softmax activation to predict dialect labels. The model is trained using categorical cross-entropy loss and the Adam optimizer. Early stopping based on validation loss is employed to prevent overfitting, with the best-performing model checkpoint restored.

A speaker-independent data split was employed. For each dialect, 10% of speakers were assigned to the validation set and 10% to the test set, with the remaining speakers used for training. Crucially, each speaker appears in exactly one of the subsets (training, validation, or test), preventing the model from exploiting speaker-specific characteristics and ensuring that performance reflects dialectal generalization rather than speaker recognition.

The complete training pipeline, including preprocessing, augmentation, embedding extraction, and model training, is implemented in a publicly available codebase[3].

4. Overall Classification Performance

Across 250 evaluation runs, the model achieved a mean and median weighted F1-score of 0.42, with a standard deviation of 0.036. These results are based on 15,397 original segments and an additional 46,272 augmented samples. Multiple runs were performed because no single fixed train–validation–test split was considered representative for the present analyses. Instead, repeated speaker-independent splits with randomly assigned

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

speakers were used to reduce split-specific bias and to ensure that analyses were not driven by particular test speakers. Importantly, all analyses focus on relative confusion patterns aggregated across evaluation runs. The number of runs was chosen empirically, as the overall results stabilized well before 250 iterations.

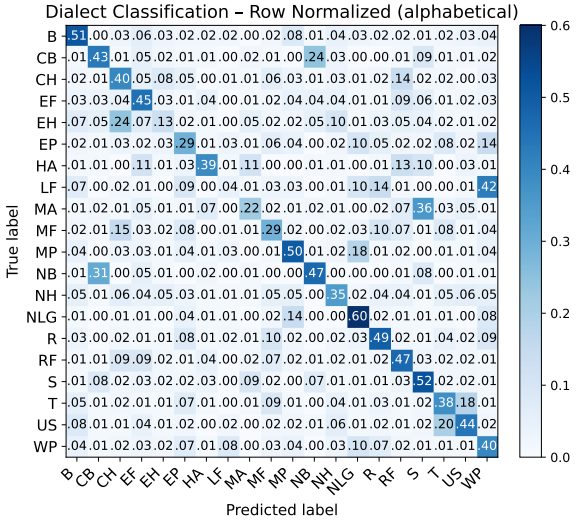

Figure 4: Row-normalized confusion matrix of dialect classification results. Abbreviations: B=Brandenburgian, CB=Central Bavarian, CH=Central Hessian, EF=East Franconian, EH=East Hessian, EP=Eastphalian, HA=High Alemannic, LF=Low Franconian, MA=Middle Alemannic, MF=Moselle Franconian, MP=Meklenburgish Pomeranian, NB=North Bavarian, NH=North Hessian, NLG=Northern Low German, R=Ripuarian, RF=Rhine Franconian, S=Swabian, T=Thuringian, US=Upper Saxon, WP=Westphalian

Figure 4 shows the row-normalized confusion matrix aggregated across all evaluation runs. The matrix indicates systematic asymmetries in misclassification behavior, suggesting that certain dialect pairs are more frequently confused than others. In addition, several dialects show low recall values, indicating that they are rarely predicted correctly.

A detailed analysis of these effects, including generational influences and systematic confusion patterns between dialects, is presented in Section 5 and Section 6.

5. Analysis of Age Groups

Figure 5 shows self-assessed speaking proficiency by birth-year group, based on the questionnaire described in Section 2.1. The birth-year cohorts can be grouped into three generations (older, middle-aged, and younger; cf. gray separators in the figure). Speakers in the older generation (three oldest cohorts) show lower variance and rate their dialect competence about two points higher (on a six-point scale) than the middle-aged generation. At the same time, a clear decrease in self-assessed dialect competence from the middle-aged to the younger generation is visible. Due to the absence of speakers born between 1970–1979, values for this cohort were interpolated solely to provide a continuous visual representation of the generational trend. These patterns in self-assessed competence are reflected in the classification results.

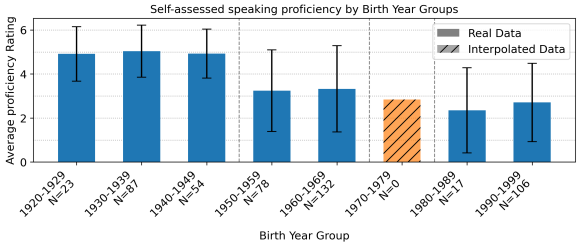

Figure 5: Average self-assessed dialect proficiency by birth-year group. Blue bars show observed data, hatched bars interpolated values, and error bars standard deviation. Gray vertical lines separate the three generations (old, middle-aged, and young).

Classification performance differs markedly across generations. The older group achieves the highest correct classification rate (46.8%), followed by the middle-aged group (45.7%), while the younger group performs markedly worse (28.1%), mirroring the lower self-assessed competence observed above.

The observed trend is consistent with established findings in dialectology. Niebaum and Macha (2014) note that dialect competence is typically correlated with speaker age, arguing that older speakers are more likely to possess dialect competence, use dialect forms more frequently, and produce more distinct dialectal realizations. Similar conclusions are drawn by Lameli (2025), who interprets this pattern as evidence of a diachronic decline in dialect competence and describes it as an intergenerational decrease in dialectal intensity.

To further isolate generational effects, separate classification experiments were conducted for each generation group, using only speakers from that generation for training, validation, and testing. Under this constraint, correct classification rates drop to 21.3% for the younger generation, 40.5% for the middle-aged generation and 39.1% for the older generation. This decrease indicates that all generations benefit from the inclusion of speech data from other age groups. While part of this effect is likely attributable to the increased amount of training data, it is also consistent with the assumption that dialectal features are, to some extent, shared

across generations despite differences in competence and usage frequency.

To further disentangle the role of shared representations from data quantity effects observed in the previous experiment, we conducted a controlled cross-generational comparison. In the baseline setting, training, validation, and testing were all performed on younger speakers. In the cross-generational setting, training and validation were performed on older speakers, while testing was performed on younger speakers. In both settings, the same set of dialects was used. For each dialect, the number of speakers was kept constant across training, validation, and test splits, and matched between the two experimental settings. Under these conditions, classification accuracy drops from 23.5% in the within-generation setting to 15.8% in the cross-generational setting. Note that the within-generation result differs from the previously reported value (21.3%) due to the restriction to a subset of dialects shared across both generations, resulting in a slightly simplified classification task. Despite this substantial decrease, performance remains clearly above chance level, indicating that dialectal features are, to some extent, shared across generations. At the same time, the performance gap suggests systematic differences in the realization of these features.

Taken together, both classification results and self-assessment data consistently indicate a generational decline in dialect competence. While younger speakers still share core dialectal characteristics with older generations, reduced dialect proficiency and weaker realization of dialect-specific features appear to limit classification performance for this group.

6. Dialect Confusion Analysis

While the overall classification performance provides a global assessment of model behavior, it does not explain why certain dialects are systematically misclassified or confused with others. In this section, we therefore analyze confusion patterns at the dialect level in order to identify linguistic, geographic, and data-related factors that may contribute to classification errors.

The four explanatory factors considered in this analysis are the number of speakers, geographical proximity, dialectal similarity, and speaker self-assessment. The number of speakers refers to the number of unique speakers per dialect included in the dataset. Each factor is represented categorically according to the levels defined in Table 2. A check mark (✓) indicates the strongest or most direct connection between two dialects, while a circle (◐) represents intermediate conditions — for example, dialects that are geographically close but separated by a transition zone, or dialects that share only partial similarity as identified in related studies. A cross (✗) denotes the weakest connection or absence of a relationship along the respective factor dimension. The classification thresholds (e.g., speaker count ranges or percentile cutoffs) are explained in the respective analytical sections.

	Factor levels		
Factor	✓	◐	✗
Speakers	<15	15-20	>20
Geo Proximity	adjacent	close	distant
Similarity	proven	partial	none
Self-assess.	<20th perc.	20th perc.-avg	>avg

Table 2: Factor coding scheme used to categorize explanatory variables. The specific thresholds and definitions for each factor are detailed in the corresponding sections of the analysis.

To structure the analysis, we examine two types of cases: (i) dialect pairs that exhibit high confusion rates (>20%), and (ii) dialects with low correct classification performance (<40%). These cases are summarized in Table 3 and form the basis for our investigation of factors that may explain the observed patterns. For confused dialect pairs, the table lists the true and predicted dialects together with the proportion of instances in which the true dialect was predicted as the other. For dialects with low correct classification performance, only the affected dialect is listed, and the reported value corresponds to its correct classification rate. In both cases, the table provides a factor-based coding of potential explanatory variables.

6.1. Number of Speakers

Figure 6 illustrates the relationship between dialect-wise classification performance and the number of speakers per dialect. Classification performance increases with the number of speakers, showing a strong positive correlation between the logarithm of speaker count and the F1-score (Pearson $\rho = 0.86$, $p < 0.001$), which is higher than the corresponding correlation in linear space (Pearson $\rho = 0.79$, $p < 0.001$). A simple linear regression model on log-transformed speaker counts yields an R^2 score of 0.73, indicating that a substantial proportion of performance variance can be explained by speaker diversity alone.

Across dialects, the number of speakers and the total number of segments per dialect are almost perfectly correlated (Spearman $\rho = 0.99$, $p < 0.001$). Nevertheless, previous experiments using voice conversion indicate that classification performance is more strongly influenced by speaker diversity than by the sheer number of segments. This moti-

	True	Predicted	Rate	Reason			
				Spk	Geo	Sim	Self
Confused	Middle Alemannic	Swabian	0.36	✓	✓	◐ Lameli (2013)	✗
	Central Bavarian Northern Bavarian	Northern Bavarian Central Bavarian	0.24 0.24	✗ ✗	✓	◐ Lameli (2013)	✗ ✗
	Low Franconian	Westphalian	0.42	✓	✓	✗	◐
	Upper Saxon Thuringian	Thuringian Upper Saxon	0.20 0.20	◐ ✗	◐	✓ Purschke (2011)	✗ ◐
	East Hessian	Central Hessian	0.24	✓	✓	✗	✗
Low recall	Eastphalian	-	0.29	✗	-	-	✓
	North Hessian	-	0.35	◐	-	-	◐
	Moselle Franconian	-	0.29	✗	-	-	◐

Table 3: Severe confusions (>20%) and low correct classification (<40%). Rate refers to either confusion rate or to correct classification rate. Reasons are evidence indicators: Spk = number of speakers, Geo = geographical proximity, Sim = dialectal similarity, Self = speaker self-assessment.

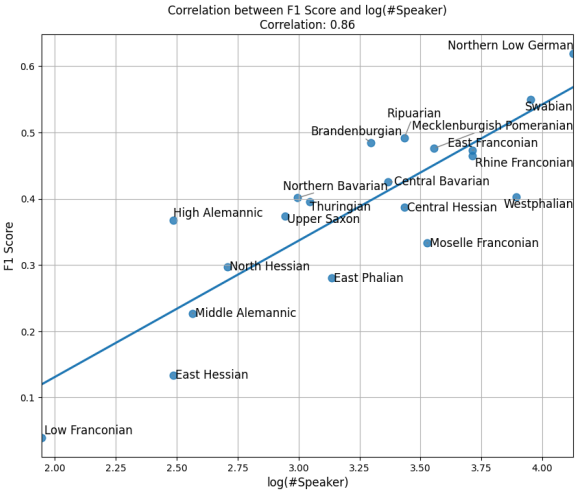

Figure 6: Relationship between the F1-score and the logarithm of the number of speakers per dialect.

vates the focus on the number of speakers rather on the number of samples.

To further characterize this relationship, we estimate a saturation point at approximately 57 speakers, determined by a slope threshold of 0.001. Beyond this point, additional speakers yield only marginal F1-score gain, suggesting diminishing returns of speaker diversity for this model.

Two additional thresholds are derived from the distribution of predicted dialect-wise F1-scores. A threshold at 0.329, corresponding to the 33.33rd percentile, is crossed at approximately 20.3 speakers, while a threshold at 0.2639 (20th percentile) is crossed at approximately 14.6 speakers. In practical terms, dialects represented by fewer than about 20 speakers are likely to fall within the lower third of classification performance, whereas dialects with fewer than roughly 15 speakers are expected to be among the lowest-performing fifth, as for East Hessian and Low Franconian. These thresholds directly motivate the speaker-based categorization used in Table 3.

6.2. Dialectal Similarity

Dialectal similarity provides a plausible explanation for several of the systematic confusions observed in the classification results. In particular, the frequent mutual confusion between Thuringian and Upper Saxon is consistent with findings reported by Purschke (2011, pp. 292f.). Purschke shows that listeners do not exhibit a clear or stable distinction between speakers of Thuringian and Upper Saxon. This pattern suggests the presence of a higher-level regional variety rather than two clearly separable dialects (see Lameli (2013)).

For the dialect pair North Bavarian and Central Bavarian, similarly high confusion rates are observed. In this case, however, the number of speakers differs more substantially (20 speakers for North Bavarian and 29 for Central Bavarian) than for the Thuringian–Upper Saxon pair (19 and 21 speakers, respectively). This imbalance may contribute to the asymmetric confusion pattern, in which North Bavarian is more frequently classified as Central Bavarian than vice versa. Both varieties of Bavarian are grouped within a broader Bavarian dialect region in regional speech classifications (Lameli, 2013, p. 194), which supports the assumption of dialectal similarity.

A comparable situation can be observed for Middle Alemannic and Swabian. According to Lameli (2013), both varieties are situated within the same

larger Alemannic dialect area. However, quantitative distance measures indicate that Swabian shows a considerable degree of structural distinctiveness within this continuum (Lameli, 2014). Consequently, dialect similarity alone is unlikely to be the primary driver of their frequent confusion.

6.3. Speaker Self-Assessment

Figure 7 shows the distribution of self-assessed speaking proficiency across dialects. Self-assessment scores range from 0 ("not at all") to 6 ("perfect") and reflect speakers' perceived active competence in their local dialect.

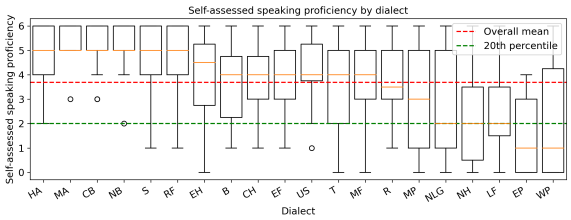

Figure 7: Distribution of self-assessed speaking proficiency by dialect. The red dashed line marks the overall mean, the green dashed line the 20th percentile. Abbreviations as in Figure 4.

A statistically significant correlation is observed between self-assessed speaking proficiency and classification performance (Spearman $\rho = 0.28$, $p < 0.001$). Although this correlation is not strong, its high statistical significance indicates that speakers who perceive themselves as more competent in their dialect tend to be classified more accurately by the model. Importantly, speaker self-assessment does not constitute an objective ground truth for dialect proficiency. Nevertheless, the observed correlation suggests that variation in dialect competence contributes to classification performance and therefore represents a relevant sociolinguistic factor.

The red dashed line in Figure 7 represents the overall mean self-assessment score, while the green dashed line marks the 20th percentile, indicating the threshold below which 20% of all responses fall. This percentile is used as a reference point for identifying dialects with comparatively low self-assessed competence.

For several frequently confused dialect pairs, including Central Bavarian–North Bavarian and Upper Saxon–Thuringian, self-assessment scores are at or above the overall average. This suggests that the observed confusions are unlikely to be primarily driven by low dialect competence, but rather by dialectal similarity. In contrast, Low Franconian, which is often confused with Westphalian, exhibits self-assessment scores slightly above the 20th percentile. While reduced dialect competence among Low Franconian speakers may contribute to confusion in this case, it is unlikely to be the dominant factor. A markedly different pattern is observed for Eastphalian. This dialect shows the lowest self-assessed speaking proficiency among all dialects in the dataset. The consistently low self-assessment scores strongly suggest that limited active dialect competence among speakers is a major contributing factor to the poor classification performance observed for Eastphalian.

6.4. Geographical Proximity

Across all frequently confused dialect pairs in Table 3, the true and predicted dialects are either adjacent or separated only by a transition area.

The only case in which the dialects are not strictly adjacent is the pair Upper Saxon and Thuringian. However, these varieties are separated merely by a transition zone and have already been shown to exhibit a high degree of dialectal similarity (cf. Section 6.2). Consequently, this case does not contradict the general pattern of spatial closeness among confused dialects.

Overall, the consistent spatial proximity of confused dialects suggests that the model captures geographically grounded variation patterns. Neighboring dialects tend to share phonetic and prosodic characteristics, which likely results in similar embedding representations and, in turn, higher confusion rates. Nevertheless, geographical proximity alone is insufficient to explain the observed patterns, as spatial adjacency does not necessarily imply linguistic similarity. The effect of geographical proximity is therefore interpreted as supportive rather than primary and is closely intertwined with dialectal similarity.

6.5. Summary

The analysis reveals that misclassifications are not random, but follow systematic patterns that can largely be explained by a combination of speaker distribution, dialectal similarity, geographical proximity, and speaker self-assessment.

Dialectal similarity emerges as a primary driver of bidirectional confusion. Dialect pairs that are confused in both directions, such as Central Bavarian–North Bavarian and Upper Saxon–Thuringian, are also documented as structurally close in the dialectology literature. These findings indicate that some dialect distinctions may be inherently difficult to model, even with increased amounts of data, and raise the possibility that such varieties could be treated as part of broader dialect regions.

The number of speakers per dialect plays a substantial role in asymmetric confusion patterns. Dialects with fewer speakers are frequently classified

as closely related dialects with larger speaker populations. This effect suggests that speaker diversity strongly influences the robustness of learned dialect representations. However, it remains unclear whether increasing the number of speakers for underrepresented dialects would fully resolve these confusions or whether strong dialectal similarity would continue to limit separability.

Speaker self-assessment shows a statistically significant but comparatively weak association with classification performance. In most confusion cases, speakers rate their dialect competence at or above the overall average, indicating that low self-assessed proficiency is not the primary cause of confusion. An exception is Eastphalian, where consistently low self-assessment scores coincide with particularly poor classification performance, suggesting limited active dialect competence as a key contributing factor.

An important exception is Moselle Franconian, which exhibits a particularly low recognition rate despite a comparatively large number of speakers and average self-assessment scores. In this case none of our tested factors sufficiently explain the observed confusion patterns, suggesting that additional factors may play a role. This is consistent with its characterization as a gradient, non-categorical variety within the Rhenish fan (Herrgen et al., 2019).

Overall, the analysis suggests that dialect classification performance is shaped by an interaction of linguistic similarity and data-related factors rather than by any single variable alone. While increasing speaker diversity may improve performance for some dialects, certain confusions appear to reflect genuine dialectal closeness that may remain challenging to resolve even with additional data.

These findings motivate a closer examination of how the model behaves in dialect transition areas, where linguistic boundaries are inherently gradient.

7. Beyond Classification: Predicting Transition Areas

Beyond discrete dialect classification, the model can be used to analyze dialect transition areas by examining the distribution of predicted dialect labels across geographical locations. Figure 8 visualizes this behavior by aggregating classification results at the level of recording locations. Each dialect is represented by a distinct color, while pie charts illustrate how often audio samples from a given location are assigned to the respective dialects. North Bavarian and Central Bavarian are here combined into a single dialect region, as well as Upper Saxon and Thuringian, based on the findings in Section 6.

To enable a meaningful analysis of transition areas, a confidence threshold was introduced to identify audio segments that cannot be reliably assigned to a single dialect. Segments that do not exceed this threshold are marked as “not properly classified” and are visualized in white. The threshold was selected based on the relationship between the proportion of unclassified segments and speakers’ self-assessed dialect proficiency. Strong negative correlations were observed for threshold values in the range between 0.25 and 0.40, indicating that higher self-assessed proficiency is associated with fewer unclassified segments in this range. A threshold of 0.40 was chosen as a compromise at the upper end of this range, ensuring a relatively strict confidence requirement while preserving the negative association.

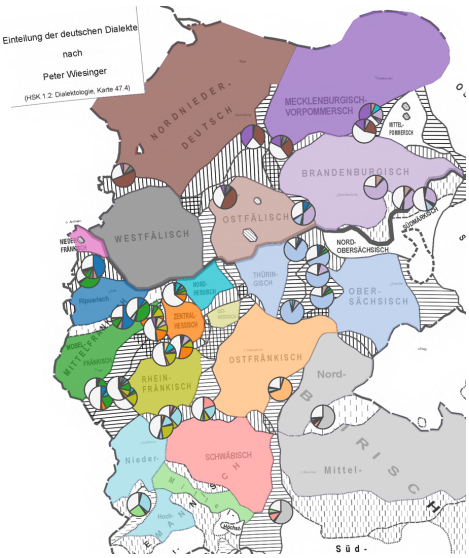

Figure 8: Predicted dialect distributions in transition areas. Colors denote dialects; pie charts show assignment proportions per location. White segments indicate unclassified samples below the confidence threshold.

Across most transition areas, the model predominantly assigns audio samples to the geographically adjacent dialect regions. Locations associated with dialects that already exhibit high confusion rates during training—such as Central Hessian, East Hessian, and Moselle Franconian—show particularly large proportions of unclassified segments. This suggests that the model’s uncertainty in transition areas mirrors the difficulty of distinguishing these dialects more generally.

The transition maps further allow for generational comparisons. Figure 9 compares predictions for the younger and the older generation only. Consistent with the generational analysis in Section 5, the younger generation exhibits a larger proportion of segments that are not properly classified. When comparing the two generations, shifts in dominant dialect assignments can be observed in selected regions, particularly in the southern area shown in

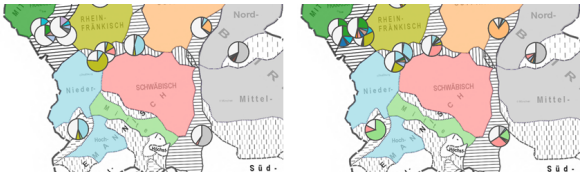

Figure 9: Excerpt from Figure 8 showing selected transition areas. The left panel displays predictions for the younger generation, while the right panel shows predictions for the older generation.

the enlarged excerpt. As illustrative examples, we consider the southern locations in the lower left and the lower right of the maps in both panels (younger and older). At these locations, which are associated with the High- and Middle Alemannic areas on the one hand and the Swabian and Bavarian areas on the other, the predicted dialect proportions differ between the younger and the older generation. This difference can also be interpreted as language change, which has already been noted in other research (Lameli (2013, p. 314)).

Taken together, these results demonstrate that the proposed model is not limited to categorical dialect classification but can also be used to explore gradient dialectal variation and generational change in transition areas.

8. Discussion

The results of this study indicate that low classification performance for certain dialects is not primarily caused by model limitations but reflects fundamental properties of dialect data. As also noted by Jokisch and Dobbriner (2019), closely related dialects can be difficult to distinguish even for human listeners, suggesting that systematic confusions often arise from inherent linguistic similarity rather than insufficient modeling.

More generally, dialects cannot be treated as strictly discrete categories. Speakers do not necessarily realize all features of a single dialect and may simultaneously use features from neighboring varieties. Dialects are therefore better modeled as fuzzy sets rather than hard classes (Ferragne and Pellegrino, 2007). This explains why certain samples remain difficult to classify with high confidence.

Generational analyses show that younger speakers achieve lower classification performance and produce more low-confidence or unclassified segments. At the same time, cross-generational training improves performance for all age groups, indicating that core dialectal features remain shared across generations despite a gradual attenuation among younger speakers.

Speaker diversity is a central aspect of dialect classification. Classification performance correlates strongly with the number of speakers, which in turn is strongly correlated with the number of samples. Nevertheless, speaker diversity is particularly important, as increasing the amount of data from only few speakers risks reinforcing speaker-specific patterns rather than dialectal variation. Unequal speaker distributions thus constitute a structural challenge, consistent with observations by Jokisch and Dobbriner (2019) and findings on diminishing returns reported by Elleuch et al. (2025).

Finally, perfect classification accuracy cannot be expected for dialect data. Not all speech samples are dialectal and assigning speakers to dialects based on place of origin does not guarantee exclusive dialect use. Moreover, as also noted by Jokisch and Dobbriner (2019), direct comparisons across studies remain difficult due to differences in corpora, class definitions, and speaker constellations. The same limitations apply to the present work, and absolute performance values should therefore be interpreted with caution.

9. Conclusion

This work shows that automatic dialect classification should not be treated as a purely accuracy-driven task. Systematic misclassifications and model uncertainty are informative and reflect linguistic structure rather than random model failure. Consequently, dialect classification models should be evaluated not only by how well they separate classes, but by how faithfully their errors reflect linguistic structure.

A detailed error analysis identified four main factors influencing confusion and low performance: speaker diversity, dialectal similarity, geographical proximity, and speaker self-assessment. Among these, the number of speakers per dialect is the most influential model-related factor, while dialectal similarity represents an inherent linguistic limitation. The analyses confirm that closely related dialects lead to systematic confusions and low-confidence predictions, consistent with dialectological accounts of gradient variation.

Generational analyses further show that dialect classification benefits from cross-generational training. Although younger speakers exhibit reduced dialectal distinctiveness, core dialectal features remain shared across generations.

Overall, the proposed model serves not only as a classifier but also as a tool for dialectological analysis: among other applications, it can, as shown in this paper, (i) reveal dialect similarity structure from systematic confusions, (ii) support apparent-time analyses of how dialects change across generations, and (iii) characterize speakers or locations in terms of dialectal affinities (illustrated here using transition areas).

10. Acknowledgements

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 two anonymous reviewers for helpful comments.

11. Bibliographical References

Sakhia Darjaa, Róbert Sabo, Marián Trnka, Milan Rusko, and Gabriela Múcsková. 2018. Automatic recognition of slovak regional dialects. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 305–308. IEEE.

Haroun Elleuch, Salima Mdhaffar, Yannick Estève, and Fethi Bougares. 2025. Adi-20: Arabic dialect identification dataset and models. In *Interspeech 2025*, pages 2775–2779. ISCA.

Emmanuel Ferragne and François Pellegrino. 2007. Automatic dialect identification: A study of british english. In *Speaker Classification II: Selected Projects*, pages 243–257. Springer.

Lea Fischbach, Akbar Karimi, Caroline Kleen, Alfred Lameli, and Lucie Flek. 2025a. Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion. In *Interspeech 2025*, pages 2780–2784.

Lea Fischbach, Akbar Karimi, Alfred Lameli, and Lucie Flek. 2025b. Edaudio: Easy data augmentation for dialectal audio. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 363—368.

Karzan Ghafoor, Sarkhel Taher, Karwan Hama Rawf, and Ayub Abdulrahman. 2025. The improved Kurdish dialect classification using data augmentation and ANOVA-based feature selection. *ARO - The Scientific Journal of Koya University*, 13:94–103.

Evangelia Gogoulou, Timothée Lesort, Magnus Boman, and Joakim Nivre. 2024. Continual learning under language shift. In *International Conference on Text, Speech, and Dialogue*, pages 71–84. Springer.

Joachim Herrgen, Jürgen Erich Schmidt, and Robert Möller. 2019. Historisches westdeutsch/rheinisch (moselfränkisch, ripuarisch, südniederfränkisch). In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum. Ein internationales Handbuch der Sprachvariation*, volume 4 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 515–550. De Gruyter Mouton, Berlin/Boston.

Nina Hosseini-Kivanani, Christoph Schommer, and Peter Gilles. 2025. Voices of luxembourg: Tackling dialect diversity in a low-resource setting. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 143–152.

Oliver Jokisch and Johanna Dobbriner. 2019. Textindependent dialect classification in read and spontaneous speech. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 350–354.

Alfred Lameli. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*, volume 54. Walter de Gruyter.

Alfred Lameli. 2014. Distanz als raumstrukturelle eigenschaft dialektaler kontaktsituationen. In Dominique Huck, editor, *Alemannische Dialektologie: Dialekte im Kontakt*, Zeitschrift für Dialektologie und Linguistik. Beihefte, pages 67–86, 297–300. Franz Steiner Verlag, Stuttgart. Beiträge der 17. Arbeitstagung für alemannische Dialektologie, Straßburg, 26.–28.10.2011.

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 51–79. De Gruyter, Berlin and Boston.

Aynalem Tesfaye Misganaw and Sabine Roller. 2022. German dialect identification and mapping for preservation and recovery. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 65–69.

Hermann Niebaum and Jürgen Macha. 2014. *Einführung in die Dialektologie des Deutschen*. De Gruyter, Berlin, Boston.

Christoph Purschke. 2011. *Regionalsprache und Hörerurteil*. Franz Steiner Verlag.

Jürgen Erich Schmidt and Joachim Herrgen and Roland Kehrein and Alfred Lameli and Hanna Fischer. 2020–. *Regionalsprache.de: Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Forschungszentrum Deutscher Sprachatlas, Philipps-Universität Marburg. 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.

Antonio Sciarretta. 2024. Dialectometry-based classification of the central–southern italian dialects. *Journal of Linguistic Geography*, 12(1):13–23.

Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157.

Samuel Stucki and Patrik Randjelovic. 2021. Automatic detection of swiss german dialects using wav2vec. Master's thesis, Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland.

Peter Wiesinger. 1983. Die Einteilung der deutschen Dialekte. In Werner Besch, editor, *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, volume 1.2 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 807–900. Berlin/New York: de Gruyter, Berlin, New York.

12. Language Resource References

Joel Shor and Subhashini Venugopalan. 2022. *TRILLsson: Distilled Universal Paralinguistic Speech Representations [Pre-trained Model]*.