Speaker Normalization via Voice Conversion Reveals a Human–Machine Dissociation in Dialect Classification

Caroline Kleen[1,*], **Lea Fischbach**[1,*], **Akbar Karimi**[2,3], **Lucie Flek**[2,3], **Alfred Lameli**[1]
[1]Research Center Deutscher Sprachatlas, Marburg University, Germany
[2]Lamarr Institute for ML and AI, Germany
[3]b-it Center, University of Bonn, Germany
{caroline.kleen, lea.fischbach}@uni-marburg.de, {akbar.karimi, lucie.flek}@uni-bonn.de, lameli@uni-marburg.de
*Equal contribution

Abstract

This study evaluates whether Retrieval-based Voice Conversion (RVC) can be used to normalize speaker-specific variability while preserving dialect-relevant acoustic cues, and what the response of human and machine systems to this manipulation reveals about the architecture of dialect recognition. In two perception experiments, speech samples from nine German dialect regions were presented either in their original form or after conversion to a single target speaker. We compared overall accuracy, confusion structures, item-level response distributions, and the interaction between listener origin and target dialect across conditions. Human classification remained stable under voice conversion. Accuracy did not differ between conditions, confusion matrices were highly correlated, and item-level divergences were minimal. The interaction between listener origin and target dialect—reflecting systematic regional bias—remained invariant. These findings indicate that RVC does not distort perceptually relevant dialectal cues and that human dialect recognition is robust to speaker normalization. In contrast, we evaluated a deep learning model under matched conditions: model accuracy improved significantly under RVC, while human performance remained unchanged. This dissociation reframes RVC as an experimental probe for investigating the divergence between human and machine speech processing, suggesting that this divergence is rooted in fundamentally different representational architectures.

Keywords: Retrieval-based voice conversion, speaker normalization, dialect classification, human perception, deep learning

1. Introduction

A defining feature of human speech perception is its robustness to inter-speaker variability. Listeners effortlessly recognize the same linguistic categories across speakers differing in age, sex, regional background, and vocal tract morphology. This is a capacity referred to as *speaker normalization* (Johnson and Sjerps, 2021). Crucially, this normalization does not operate on absolute acoustic values but encodes linguistic categories relative to speaker-specific context (Sjerps et al., 2019). Dialect recognition represents a particularly demanding instance of this capacity. Listeners must simultaneously normalize for speaker-level variability and attend to dialect-level variation, which is itself encoded in the same acoustic dimensions that differ across speakers.

Deep learning models for dialect classification face an analogous challenge, but address it through an entirely different mechanism. Rather than deploying pre-structured normalization routines shaped by years of linguistic exposure, neural models optimize directly on the statistical structure of the training data. As a consequence, their internal representations may conflate speaker-specific and dialect-specific variance whenever these co-occur in the training signal. This problem is well-known in low-resource settings where limited speaker diversity can lead models to overfit to speaker characteristics rather than linguistic variation (Fischbach et al., 2025a).

Retrieval-based Voice Conversion (RVC) offers a way to disentangle these sources of variance computationally. By mapping utterances from multiple speakers onto a single target voice, RVC attenuates speaker-specific acoustic properties while preserving the linguistic signal (Sisman et al., 2020). Previous work has shown that this procedure improves dialect classification performance in low-resource settings (Fischbach et al., 2025a) and geographic regression tasks (Gutscher and Pucher, 2025). However, these findings address the technical validity of RVC as an augmentation method. It remains unclear whether RVC constitutes a valid normalization technique for dialectal speech resources beyond performance gains. A different question is what RVC reveals about the architecture of dialect recognition itself. These two questions motivate the present study.

We address this question by adopting a parallel evaluation design in which the same stimulus material—original and RVC-converted dialectal speech from nine German dialect regions—is presented to both human listeners and a neural classifier under matched task conditions. We hy-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 177–187
16 May 2026.

pothesize that if human and machine dialect recognition rely on the same acoustic representations, both should respond similarly to RVC. If they rely on different representations, their responses should diverge. We exploit this logic to investigate whether speaker normalization by RVC is compatible with the cognitive normalization mechanisms that human listeners deploy. The results reveal a sharp dissociation. RVC systematically improves machine classification while leaving human performance entirely unaffected. We argue that this dissociation is not incidental but reflects a fundamental difference in how biological and artificial systems represent dialectal speech.

The remainder of the paper is organized as follows. Section 2 presents the material and methods: corpus and stimuli, the acoustic characterization of RVC, and the evaluation design including both the computational and perceptual experiments. Section 3 presents the results for human and model classification and quantifies the dissociation directly. Section 4 discusses theoretical implications, and Section 5 concludes.

2. Material and Methods

2.1. Data

The stimuli presented to human listeners are randomly sampled from the same pool of recordings used for model training and evaluation (the REDE corpus (Schmidt et al., 2020–)), supporting comparability across conditions, although the exact stimulus sets differ between the two evaluation setups. The REDE corpus comprises recordings from male speakers[1], from three age groups throughout Germany: young (17–26 years), middle-aged (42–59 years) and older speakers (60+ years), captured in five to six recording situations[2]. For the purpose of this study, only data from the older age group is used, as this group has been shown to exhibit particularly high dialect competence (Lameli, 2025), providing stimuli that maximize the signal-to-noise ratio for dialect-relevant acoustic features, which is a prerequisite for a clean experimental test of normalization robustness. Furthermore, only the recording situation of the so-called dialectal "Wenker-Sentences" (Wenker, 2013) are analyzed. In this task, the participants translated 40 standard German sentences, spoken by the interviewer, into their local dialect.

The recordings are first automatically segmented using a speaker diarization pipeline to identify stretches of speech belonging to the target speaker; details of this procedure are reported in a companion paper (Fischbach, 2024). These excerpts are subsequently concatenated per speaker and divided into consecutive 10-second segments to obtain multiple samples per speaker of equal duration. The choice of 10 seconds is motivated by preliminary experiments indicating that shorter durations lead to noticeably lower classification performance by a Deep Learning (DL)-model, whereas longer segments did not result in significant performance gains. Ten-second segments are also a practical duration for perceptual evaluation: they provide sufficient phonological material for dialect recognition while keeping listening fatigue within acceptable bounds for an online perception study. All audio files are pre-processed by converting them to mono, setting the bit depth to 16 bits, and the sampling rate to 16 kHz, which matches with the specifications of the DL-model used for this experiment.

The classification tasks consider a total of ten distinct German dialect areas, categorized according to Lameli (2013). Due to insufficient data, the Frisian language area was excluded from the experiments, resulting in nine dialect classes. Figure 1 illustrates the spatial distribution of these nine dialect areas across Germany; the excluded Frisian area is located in the northwestern corner of the map.

Figure 1: Dialect classification of Germany according to Lameli (2013), serving as basis for the classification tasks.

Table 1 provides an overview of these dialect areas, including the number of 10-second samples and speakers used for both the DL-experiment and the human perception study. The first two columns report the total number of generated segments and the corresponding number of speakers in the DL setup, where all available excerpts are included, resulting in a naturally imbalanced

[1]Only male speakers are included, as the REDE project focused on this group for methodological consistency.

[2]Further specifications can be found here: https://rede-infothek.dsa.info/

class distribution across dialect areas. The third column shows the subset selected for the human study. As presenting the full set of segments would have been impractical for perceptual evaluation, the study was restricted to 100 samples in total. The relative class proportions of the DL dataset were retained to ensure comparability between model-based and human evaluation. Within each dialect area, only one sample per speaker was selected in order to maximize speaker diversity. Consequently, in the human study, the number of samples equals the number of speakers per dialect area. In total, the dataset comprises recordings from 166 speakers, amounting to 14.36h of audio data. The recordings originate from 91 different locations in Germany, corresponding to an average of approximately 18.44 speakers per dialect class.

Dialect area	#Samples (DL)	#Speaker (DL)	#Samples (Study)
Bavarian	387	15	8
Brandenburgish	167	4	3
East Central German	462	14	9
East Franconian	512	12	9
Northern Low German	916	34	19
Southern Low German	733	22	14
West Central German	847	30	17
West German	348	10	6
West Upper German	798	25	15
Total	5170	166	100

Table 1: Distribution of 10-second speech segments across dialect areas for the deep learning (DL) experiment and the human perception study (Study). The DL setup includes all available segments. For the perceptual study, a proportional subset of 100 segments was selected. In the human study, the number of samples equals the number of speakers per dialect area.

2.2. Retrieval-based Voice Conversion

Voice conversion is performed using the widely adopted RVCv2 model[3], which standardizes speaker-related acoustic cues by converting all samples to the voice of a single target speaker. As target voice, we employ the publicly available pre-trained weights corresponding to the Jonathan Frakes (German "Beyond Belief: Fact or Fiction" Mix) RVC voice model[4], where *Mix* refers to the use of recordings from multiple German dubbing actors during training. This target voice acoustically corresponds to a male adult speaker of middle age. Previous work has shown that the choice of target speaker, including differences in age groups, does not significantly affect performance when using RVC-based audios for dialect classification, even when multiple or age-matched target voices are employed (Fischbach et al., 2025a).

To evaluate the acoustic impact of the voice conversion process, we analyzed three key acoustic features before and after applying Retrieval-based Voice Conversion (RVC): mean pitch, articulation rate (AR) and the first three formant frequencies (F1, F2, F3). All acoustic features were extracted using Praat (Boersma and Weenink, 2022), executed via Parselmouth (Jadoul et al., 2018) in Python, including the use of the VocalToolkit plugin (Corretge, 2012–2024) for the AR. Since all acoustic features are time-varying, mean values were computed per 10-second segment. The final reported values reflect the mean and standard deviation of these segment-level means. The mean pitch shows only a minor change between the original (118.55 ± 10.71 Hz) and RVC version (118.10 ± 10.68 Hz), with a nearly identical pitch range (59.72 vs. 59.71 Hz). The AR remains similarly stable (original: 4.39 ± 0.49 syl/sec; RVC: 4.42 ± 0.46 syl/sec), indicating preserved intonational and temporal structure. In contrast, the long-term average formant (LTAF) frequencies are more affected, although both the difference between the original and RVC and the variation range (see error bars in Figure 2) are negligible in phonetical and phonological terms. Also, the variation is within the allophonic range, as evidenced by Sendlmeier and Seebode (2006) for the vowels of German: F1 increases slightly (from 536.20 ± 53.38 Hz to 625.22 ± 48.06 Hz), F2 decreases slightly (from 1625.51 ± 89.27 Hz to 1597.58 ± 72.89 Hz), and F3 remains relatively stable (from 2616.68 ± 92.17 Hz to 2667.59 ± 41.75 Hz).

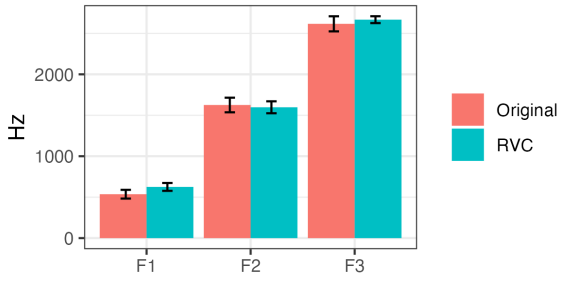


Figure 2: Mean long-term average formant frequencies (F1–F3) before and after voice conversion. Bars represent the mean of segment-level averages (10 s), and error bars indicate the corresponding standard deviation across segments.

These acoustic findings carry different implications for human and machine processing. The sta-

[3] https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI

[4] https://huggingface.co/Connum/RVC-models/resolve/main/frakes_xfactor_ger.zip

bility of F0 and AR suggests that suprasegmental features relevant to both dialect perception and model classification are preserved. The shifts in F1–F3, by contrast, indicate that RVC primarily targets the spectral envelope, which is the primary acoustic correlate of perceived vocal identity (Fant, 1971). For human listeners, this transformation may be largely inconsequential: research on talker normalization suggests that listeners encode linguistic categories relative to speaker-specific context rather than in absolute acoustic terms (Sjerps et al., 2019), implying that a shift in the overall formant space should not distort within-speaker contrasts that carry dialectal information. For a neural model, however, formant-level shifts may have more complex consequences: they reduce between-speaker variance in the training signal, potentially making dialect-specific patterns more salient. This asymmetry generates a testable prediction that RVC should benefit model performance while leaving human performance unaffected. The following evaluation—which presents comparable stimuli to both human listeners and a neural classifier—directly tests this prediction by comparing the response of each system to the same acoustic manipulation.

2.3. Experiments

To test the predicted dissociation between human and machine dialect recognition under RVC, we adopt a parallel evaluation design in which both systems are exposed to stimuli drawn from the same underlying corpus and perform the same dialect classification task. Rather than a strictly item-matched setup, the two evaluations are designed to be comparable at the level of task, label space, and data distribution, while respecting the methodological constraints of each system.

Specifically, the human study is based on a fixed subset of 100 stimuli (one per speaker) to avoid speaker repetition and perceptual learning effects, whereas the model is evaluated on the full dataset using repeated speaker-disjunct train–validation–test splits. This ensures that the model does not memorize speaker-specific characteristics and provides a robust estimate of generalization performance across the data distribution.

As a result, the comparison should be interpreted as a parallel evaluation of two systems under comparable task conditions, rather than a strictly item-by-item comparison. This design allows us to attribute differences in performance patterns to differences in processing strategies, while accounting for the distinct constraints of human perception and machine learning.

2.3.1. RVC-Augmentation for Deep Learning

The audio material described above serves as input for Google's TRILLsson model (Shor and Venugopalan, 2022)[5], which extracts a vector representation (embedding) for each audio segment. These embeddings are then used to train and evaluate a lightweight Multilayer Perceptron (MLP) consisting of two hidden layers and an output layer with 297,876 trainable parameters. To prevent the model from memorizing speaker-specific characteristics, we apply a strict speaker-disjunct train–validation–test split, ensuring that each speaker appears in only one partition. For each dialect D with $|S_D|$ speakers, we randomly select $\lceil |S_D|/10 \rceil$ speakers for the validation set and $\lceil |S_D|/10 \rceil$ speakers for the test set (the remaining speakers form the training set). Because speaker selection directly affects performance and to get a more reliable estimate of the model's real-world performance, we repeat this procedure across 250 independent runs with different randomly sampled subsets for each run[6].

In the baseline condition, the model is trained exclusively on the original recordings. In the augmented condition, RVC-converted versions of the training samples are added to the training data, while the speaker-disjunct partitioning is maintained. Without RVC-augmentation, the MLP achieves a classification accuracy—defined as the proportion of correctly predicted samples—of 52.7% ($SD = 7.4\%$), whereas with RVC-augmentation the accuracy is increased to 57.2% ($SD = 6.7\%$). This improvement is statistically significant, $t(498) = 7.26$, $p < .001$, $d = 0.65$, based on comparing the 250 baseline runs to the 250 RVC-augmented runs as independent samples. This condition serves as the machine-side component of the parallel evaluation, providing a performance baseline against which the predicted benefit of RVC can be assessed.

2.3.2. RVC in Human Perception

The human perception studies constitute the second component of the parallel evaluation. Critically, the stimuli presented to listeners are drawn from the same pool as those used for model training and testing, ensuring direct comparability of conditions. To contextualize the classification performance of the model presented above, we conducted two perception studies with linguistic laypeople (Anders, 2010),

[5] TRILLsson was selected because it was developed for paralinguistic classification tasks and provides a general-purpose acoustic embedding space without dialect-specific supervision.

[6] The complete pipeline is available at: https://github.com/WoLFi22/DialectClassificationPipeline

which are defined as people without knowledge of linguistics (Klein, 2021). A between-subjects design was adopted to prevent carry-over effects: listeners who first classify original recordings might develop perceptual strategies that transfer to the converted condition, confounding the comparison.

Related work has shown that the design of such tasks strongly affects participants' performance: Hundt et al. (2015) reported accuracies of about 64% when lay participants selected from predefined location options, whereas Kleen (2022) found considerably lower classification performance of 38% in a map-based task without predefined choices. Building on these findings, our perception studies employ a format that minimizes external guidance while still allowing for meaningful geographical categorization, thereby ensuring conceptual comparability with the model, which is likewise evaluated on dialect area categories rather than fixed locations. The map shown to participants, depicting these dialect areas, is presented in Figure 1.

The first study (**Study-O**) investigates how participants classify the original recordings from the corpus, while the second study (**Study-VC**) examines the participants' ratings of the RVC-modified samples. This allows for a direct comparison between human listeners and the model. The surveys were implemented using Gorilla Experiment Builder (`www.gorilla.sc`) to create and host the experiment (Anwyl-Irvine et al., 2020) and distributed using Prolific (`www.prolific.com`) (Prolific, 2024). A total of 100 native German (L1) participants were each presented with all 100 stimuli (10 seconds each) in individually randomized order. The study thus comprised 100 trials per participant, with an estimated completion time of approximately 20–30 minutes including response time and attention checks. Using a geographical map, participants were asked to assign each stimulus to one of ten dialect classes (e.g., Bavarian), with an additional "don't know" option to avoid forced-choice effects in cases of high uncertainty. Although Frisian was included as a response category, no Frisian stimuli were presented. To ensure a direct comparison with the model's forced-choice classification, responses selecting either Frisian or "don't know" were excluded from the subsequent accuracy analyses. As shown in Table 1, the distribution of the 100 stimuli across dialects follows the distribution of the original corpus, and consequently also reflects the number of available speakers per dialect. Importantly, each stimulus in the study originates from a different speaker, ensuring that listeners never evaluated two samples from the same speaker. Participants' engagement was monitored using periodic attention checks (one after every 20 trials). In these trials, no audio stimulus was presented; instead, participants were instructed to select a specific category (e.g., "Please select Bavarian for this trial") to verify they were still attending to the task. Responses to these attention checks were used solely for participant screening and were excluded from the dialect classification analysis.

In Study-O, 50 participants (31 male, 18 female, 1 non-binary; $M_{\text{age}} = 30.06$, $SD = 5.20$ years) achieved an overall classification accuracy of 31.66% ($SD = 13.51\%$). In Study-VC, 50 participants (32 male, 17 female, 1 non-binary; $M_{\text{age}} = 30.06$, $SD = 5.02$ years) achieved an accuracy of 31.38% ($SD = 13.87\%$), statistically indistinguishable from the original condition. A two-sample t-test, based on each participant's total score, did not reveal significant differences in classification performance between the two groups, $t(98) = 0.10$, $p = .919$. A formal equivalence test is reported in Section 3.1.

Participants in both studies were geographically distributed across Germany, ensuring broad regional representation. The regional distributions were highly similar across the two participant groups.

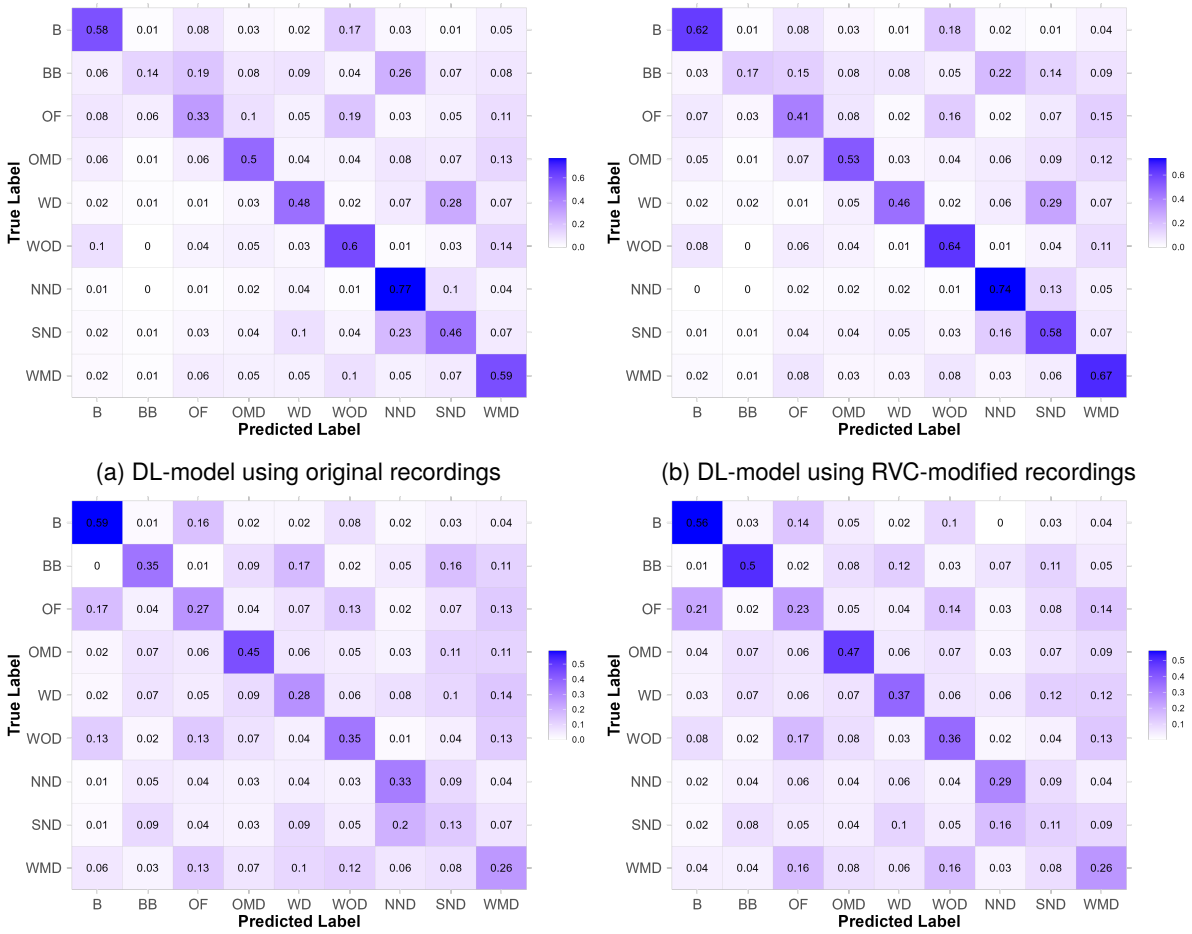
3. Results

We report results in three steps: first, the overall effect of voice conversion on human classification and the modulating role of listeners' geographical origin; second, the model results and third, the direct quantification of the human–machine dissociation.

3.1. Human Classification

3.1.1. Effect of Voice Conversion on human perception

Consistent with the prediction derived from talker normalization theory, human dialect classification showed no response to voice conversion across any of the metrics examined (Study-O and Study-VC). Listener accuracy remained stable (31.66% vs. 31.38%). A Pearson correlation of $r = 0.96$ between the confusion matrices (Figure 3) confirms that error patterns were highly consistent across conditions. We conducted a per-dialect F1-score analysis to ensure that the voice conversion preserves the phonetic dialectal markers. The human performance remains remarkably stable across conditions: while the dialect classes Brandenburgish (+0.102), West German (+0.083) and Northern Low German (+0.052) showed a slight increase in F1-scores, results for Bavarian (–0.015), East Central German (±0.00), West Central German (–0.003), Southern Low German (–0.011) and West Upper German (–0.007) were nearly identical. Only East Franconian showed a marginal decrease (–0.049). These findings confirm that



(a) DL-model using original recordings

(b) DL-model using RVC-modified recordings

(c) Humans using original recordings

(d) Humans using RVC-modified recordings

Figure 3: Confusion matrices showing classification frequencies for the DL-model (a, b) and human listeners (c, d), before and after RVC augmentation. Rows indicate true labels, columns predicted ones. Abbreviations: B = Bavarian; BB = Brandenburgish; OF = East Franconian; OMD = East Central German; WD = West German; WOD = West Upper German; NND = Northern Low German; SND = Southern Low German; WMD = West Central German.

RVC-generated stimuli remain perceptually valid and representative of the target dialects.

To assess whether this stability also holds at the level of individual items, we computed the Jensen-Shannon divergence between the response distributions for each stimulus across conditions. The divergence was consistently low ($M = .091$, $SD = .037$), indicating that the same items were perceived similarly, regardless of whether the original or converted version was presented. Together, these results demonstrate invariance across global performance, confusion structure, and item-level perception across conditions.

To move beyond the absence of a significant difference and provide positive evidence for equivalence, we conducted a two one-sided tests (TOST) equivalence analysis (Lakens, 2017), with equivalence bounds set at ± 5 percentage points, which is a margin we consider the smallest difference of practical significance given the overall accuracy range observed. The analysis confirmed statistical equivalence between the two conditions, $t(98) = -1.72, p = .044$, establishing that human performance under RVC falls within the range of the original condition not merely by failure to reject the null, but by positive equivalence criterion.

3.1.2. Effect of listeners' geographical origin

Beyond general confusion patterns, we investigated whether the normalization process affects regional perception and how classification performance is moderated by the listeners' background. As illustrated in the heatmaps (Figure 4), accuracy is systematically influenced by the geographical origin of the participants. To test the stability of this effect across conditions, we fitted a linear mixed-effects model including a triple interaction between listener origin, target dialect, and study condition (original vs. VC). The analysis revealed a significant main

effect of listener origin, $F(8, 84.3) = 2.37, p = .024$, suggesting that participants' general classification ability differs slightly depending on their own regional origin. In contrast, the main effect of study condition was not significant, $F(1, 84.3) = 0.01, p = .916$, confirming that overall accuracy remained identical between original and voice-converted stimuli. Most importantly, we found a highly significant interaction between listener origin and target dialect, $F(64, 8603.4) = 4.81, p < .001$. Crucially, the triple interaction with study condition was not significant, $F(64, 8603.4) = 1.19, p = .143$. This statistically confirms that the "regional bias" remains invariant across conditions: RVC-based speaker normalization does not alter the cognitive strategies or regional cues listeners rely on.

These results provide a robust explanation for the observed human performance: while accuracy is influenced by both the listener's origin and its interaction with the target dialect, it remains unaffected by the voice conversion process.

3.2. Model Classification

In contrast to the invariant human performance, the deep learning classifier responded systematically to voice conversion. Accuracy increased from 52.7% ($SD = 7.4\%$) to 57.2% ($SD = 6.7\%$), a statistically significant improvement, $t(498) = 7.26, p < .001, d = 0.65$. While the confusion matrices remained highly correlated across conditions ($r = 0.99$), indicating that the overall classification structure was preserved, the consistent accuracy gain across 250 independent runs rules out sampling artifacts and confirms a genuine benefit of speaker normalization for the model.

To control for the potential confound between RVC-specific effects and increased training data quantity, we conducted an additional size-matched augmentation experiment. Specifically, we applied frequency masking to each training sample, following the augmentation strategy proposed by Fischbach et al. (2025b), which was shown to be the most effective augmentation method for dialect classification in prior work. This ensures that the total number of training instances is identical to the RVC-augmented condition. This non-RVC augmentation yields a mean classification accuracy of 55.6% ($SD = 6.3\%$). A direct comparison between the frequency-masked and RVC-augmented conditions shows a statistically significant difference, $t(498) = 2.28$, $p = .023, d = 0.20$, indicating a small but reliable advantage of RVC over generic augmentation. These results suggest that while part of the performance gain can be attributed to increased training data, RVC provides an additional benefit beyond standard augmentation.

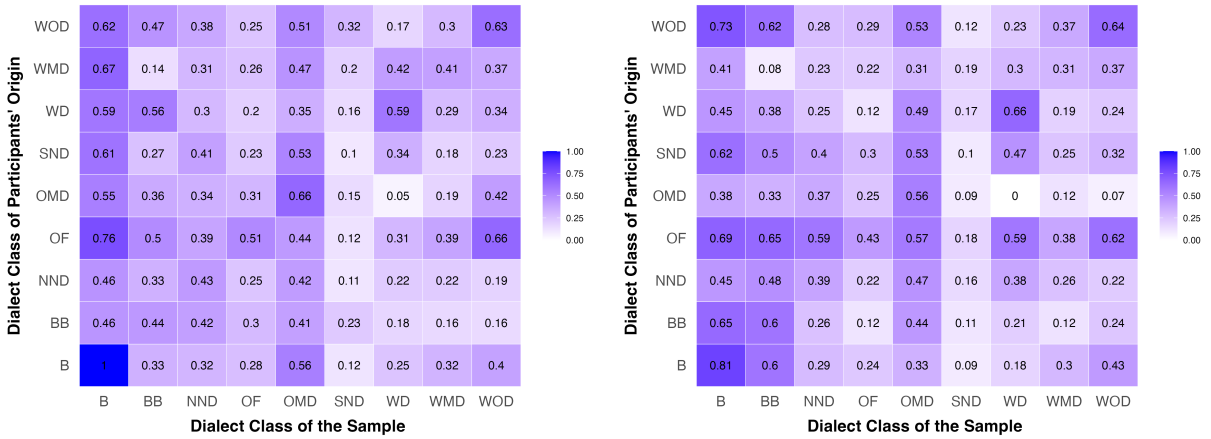
To assess whether the observed improvements are consistent across dialect classes, we additionally conducted a per-dialect analysis using one-vs-rest F1-scores, which provide a more class-sensitive evaluation under dataset imbalance than overall accuracy. The results show consistent improvements for all dialects, with gains ranging from +0.019 to +0.073. Importantly, these improvements are observed both for low-resource dialects (e.g., Brandenburgish, 4 speakers, +0.053) and high-resource dialects (e.g., West Central German, 30 speakers, +0.044), suggesting that the effect is not driven by a subset of dominant classes. Overall, this analysis indicates that the benefit of RVC generalizes across dialect classes and is not an artifact of dataset imbalance.

Notably, the confusion structure of the model differs substantially from that of human listeners even in the baseline condition (see Figures 3a and 3c), suggesting that the two systems do not rely on the same discriminative features. Section 3.3 quantifies this cross-system divergence directly.

3.3. Dissociation

Taken together, the results reveal a double dissociation between human and model behavior under voice conversion. Human accuracy remained stable (31.66% vs. 31.38%), while model accuracy improved (52.7% vs. 57.2%). To quantify this dissociation, we compare condition effect sizes across systems rather than aggregating incommensurable accuracy scales. For the model, RVC produced a gain of 4.5 percentage points, confirmed as statistically significant with a large effect ($d = 0.65, p < .001$, based on 250 independent runs). For human listeners, the corresponding difference was -0.28 percentage points, statistically indistinguishable from zero ($t(98) = 0.10, p = .919$) and confirmed as equivalent to zero by formal TOST criterion (Section 3.1.2). The contrast between the model's effect size ($d = 0.65$) and the near-zero effect for human listeners ($d = 0.02$) directly indexes the magnitude of the dissociation. That the model benefits substantially from a manipulation to which human listeners are entirely insensitive cannot be attributed to differences in task difficulty, stimulus quality, or sampling variability.

A second dimension of the dissociation concerns the confusion structure. While human confusion matrices were highly correlated across conditions ($r = 0.96$), and model confusion matrices similarly stable ($r = 0.99$), the cross-system correlation between human and model confusions was substantially lower ($r = 0.80$, computed as the Pearson correlation between corresponding cells of the two normalized confusion matrices). This indicates that humans and the model not only respond differently to RVC, but classify dialects differently in the first place. This is a finding that bears directly on the interpretation of what each system has learned.



(a) Human classification accuracy by participant origin (Study-O)

(b) Human classification accuracy by participant origin (Study-VC)

Figure 4: Heatmaps showing classification accuracy for the human listeners in Study-O (left) and Study-VC (right). Abbreviations: B = Bavarian; BB = Brandenburgish; OF = East Franconian; OMD = East Central German; WD = West German; WOD = West Upper German; NND = Northern Low German; SND = Southern Low German; WMD = West Central German.

4. Discussion

The central finding of this study is not that RVC improves model performance, nor that human dialect recognition is robust; both results are consistent with prior expectations. The theoretically consequential finding is the combination, namely that the same acoustic manipulation produces opposite responses in the two systems. This double dissociation cannot be explained by differences in task difficulty, stimulus quality, or chance. It reflects a structural difference in how dialectal speech is represented.

The performance gap between human lay listeners (31.66%) and the neural model (52.7%) reflects a fundamental difference in classification strategies. While the model leverages high-dimensional statistical regularities across the entire frequency spectrum, humans rely on a limited set of "shibboleths" or stereotypical markers (e.g., Bavarian features) shaped by media exposure and personal mobility (Köster et al., 2012; Sauer and Hoffmeister, 2022; Hundt et al., 2015). In brief 10-second segments, these prototypical markers may be difficult to detect, forcing lay listeners to rely on subjective mental maps. For human listeners, the results are consistent with exemplar-based models of speech perception (Goldinger, 1998; Johnson, 1997), in which stored representations include speaker-specific information that listeners use as a normalization baseline. Under this account, converting all speakers to the same voice does not help listeners—they already normalize implicitly—but it also does not hurt them, because the normalization baseline adjusts to the new voice. As detailed below, this account receives particularly strong support from the invariance of the regional bias across conditions. This finding goes beyond overall accuracy to the fine-grained structure of dialectal competence.

One alternative interpretation is that human accuracy (31.66%) was too variable or too far from ceiling to detect a subtle RVC effect. However, the TOST analysis rules this out by establishing positive equivalence rather than mere non-significance, and the item-level stability (JSD = 0.091) confirms that individual stimuli were not systematically revalued.

The invariance of the regional bias across conditions provides a particularly stringent test of this account. The interaction between listener origin and target dialect—reflecting the finding that listeners classify nearby dialects more accurately than distant ones (Hundt et al., 2015)—was fully preserved under RVC. This result goes beyond showing that overall accuracy is unaffected: it demonstrates that the fine-grained structure of dialectal competence, shaped by each listener's individual exposure history, is equally unaffected. Under a speaker normalization account, this is precisely what one would expect: if listeners encode dialectal categories relative to speaker context, then shifting the speaker context uniformly—as RVC does—should leave the relative structure of dialectal knowledge intact. The data confirm this prediction at the level of individual listener-dialect interactions.

For the neural classifier, the improvement under RVC is interpretable within a straightforward signal-processing framework: by reducing between-speaker variance, RVC makes dialect-specific patterns more consistently accessible to the model across training instances. This is the mechanism proposed by Fischbach et al. (2025a). What the

present study adds is the contrastive context: the model's sensitivity to speaker-level variance is not a feature of dialect classification per se, but a consequence of the model's architecture. A system with genuine normalization capacity—such as a human listener—should be indifferent to this manipulation. The model is not.

This has practical implications for the use of RVC as an augmentation technique. The human data establish that RVC does not distort dialectally relevant information, providing the perceptual validity that model improvements alone cannot guarantee. RVC-augmented data can therefore be considered dialectally equivalent to original recordings from the perspective of linguistic content. This makes RVC a methodologically validated tool for dialect data augmentation in settings where speaker diversity is limited, with direct applicability to low-resource dialect NLP pipelines.

While the present findings are interpretable within a coherent theoretical framework, they also open specific empirical questions that the current design cannot resolve. Two directions for future research follow directly from these findings. First, the present design used a between-subjects comparison; a within-subjects replication would allow direct modeling of individual-level stability and increase sensitivity for detecting subtle condition differences. Second, the formant shifts introduced by RVC (particularly $\Delta F1 \approx 90\,\mathrm{Hz}$) raise the question of whether dialect-specific vowel contrasts are preserved at the phonological level, not merely at the perceptual level. While the perceptual data suggest that these shifts do not distort dialect-relevant contrasts, a targeted analysis of vowel space geometry before and after conversion, stratified by dialect region, could provide additional phonological corroboration of this conclusion.

5. Conclusion

This study pursued two objectives: to evaluate whether Retrieval-based Voice Conversion constitutes a valid normalization technique for dialectal speech resources, and to determine what the response of human and machine systems to this manipulation reveals about the architecture of dialect recognition. The answer is affirmative, but the more consequential finding concerns what RVC reveals about the nature of dialect recognition itself. The sharp dissociation between human stability and model improvement under identical acoustic manipulation demonstrates that these two systems do not share a common representational basis for dialect classification. Human listeners deploy speaker normalization as a cognitive default, rendering their performance robust to external speaker transformations. Neural classifiers, by contrast, are sensitive to the statistical structure of the training signal and benefit when speaker-level variance is reduced externally.

This dissociation has methodological and theoretical implications that extend beyond dialect classification. Methodologically, it establishes that human perception studies and computational modeling address different questions about speech, and that improvements in one domain do not automatically transfer to the other. Theoretically, it suggests that the gap between human and machine speech processing is not merely quantitative—which is a matter of performance levels—but qualitative, rooted in fundamentally different representational architectures. For the applied question of whether RVC constitutes a valid normalization technique for dialectal speech resources, the answer is unambiguous: dialectally relevant information is preserved, and RVC-augmented data are perceptually equivalent to original recordings. For the broader question of what RVC reveals about speech processing, the answer is more consequential. The human-machine dissociation documented here suggests that the architecture of dialect recognition—biological or artificial—determines how systems respond to speaker normalization. This architectural difference cannot be read off from performance metrics alone.

6. Limitations

The exclusive use of older male speakers limits the generalizability of the findings to mixed-gender and mixed-age settings; future work should examine whether the observed dissociation holds across speaker sex and age.

7. 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.

8. Bibliographical References

Christina A. Anders. 2010. Die wahrnehmungsdialektologische Rekodierung von laienlinguistischem Alltagswissen. *Perceptual Dialectology: Neue Wege der Dialektologie*, pages 67–87.

Adam L. Anwyl-Irvine, Jessica Massonnié, Amy Flitton, Natasha Z. Kirkham, and John K. Evershed. 2020. Gorilla in our midst: an online behavioural experiment builder. *Behavior Research Methods*, 52:388–407.

Matthew Baas and Herman Kamper. 2022. Voice Conversion Can Improve ASR in Very Low-Resource Settings. In *Proc. Interspeech 2022*, pages 3513–3517.

Cynthia G Clopper and David B Pisoni. 2004. Some acoustic cues for the perceptual categorization of american english regional dialects. *journal of phonetics*, 32(1):111–140.

Gunnar Fant. 1971. *Acoustic Theory of Speech Production*. De Gruyter Mouton, Berlin, Boston.

Lea Fischbach. 2024. A comparative analysis of speaker diarization models: Creating a dataset for German dialectal speech. In *Proceedings of the Third Workshop on NLP Applications to Field Linguistics*, pages 43–51, Bangkok, Thailand. Association for Computational Linguistics.

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, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Marja Gackstatter and Oliver Niebuhr. 2013. Eine kontrastive phonetische analyse niederdeutscher langvokale. *Linguistik Online*, 53(3).

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.

Stephen D. Goldinger. 1998. Echoes of echoes? an episodic theory of lexical access. *Psychological review*, 105(2):251–279.

Lorenz Gutscher and Michael Pucher. 2025. Audio-based classification and geographic regression of austrian dialects. In *Proc. Interspeech 2025*, pages 2765–2769.

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.

Markus Hundt, Nicole Palliwoda, and Saskia Schröder. 2015. Wahrnehmungsdialektologie – Der deutsche Sprachraum aus der Sicht linguistischer Laien: Exemplarische Ergebnisse des Kieler DFG-Projekts. In Roland Kehrein, Alfred Lameli, and Stefan Rabanus, editors, *Regionale Variation des Deutschen: Projekte und Perspektiven*, pages 585–629. De Gruyter Mouton, Berlin, Boston.

Keith Johnson. 1997. Speech perception without speaker normalization: An exemplar model. In Keith Johnson and John W. Mullennix, editors, *Talker Variability in Speech Processing*, pages 145–165. Academic Press, San Diego.

Keith Johnson and Matthias J Sjerps. 2021. Speaker normalization in speech perception. *The handbook of speech perception*, pages 145–176.

Caroline Kleen. 2022. Identifikation von regionalakzenten durch linguistische laien. Master's thesis, University of Trier. Unpublished.

Wolf Peter Klein. 2021. *Was denken linguistische Laien über die (deutsche) Grammatik?: Beobachtungen und Interpretationen anhand des öffentlichen Sprachgebrauchs*, pages 227–248. De Gruyter, Berlin, Boston.

Stefan Kleiner. 2017. F1/f2-diagramme als darstellungsmittel bairisch geprägter standardsprachlicher vokalsysteme. In Alexandra N. Lenz, Ludwig Maximilian Breuer, Tim Kallenborn, Peter Ernst, Manfred Michael Glauninger, and Franz Patocka, editors, *Bayerisch-österreichische Varietäten zu Beginn des 21. Jahrhunderts: theorie- und korpusbasierte Ansätze*, volume 167 of *Zeitschrift für Dialektologie und Linguistik – Beihefte*, pages 263–284. Franz Steiner Verlag, Stuttgart.

Olaf Köster, Roland Kehrein, Karen Masthoff, and Yasmin Hadj Boubaker. 2012. The tell-tale accent: Identification of regionally marked speech in German telephone conversations by forensic phoneticians. *International Journal of Speech Language and the Law*, 19(1):51–71.

Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.

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

Alfred Lameli. 2025. *Gesprochenes Deutsch in den Regionen: Eine Standortbestimmung für die Bundesrepublik Deutschland*, pages 51–80. De Gruyter, Berlin, Boston.

Prolific. 2024. Prolific (version used: [May 2025]). Online platform.

Verena Sauer and Toke Hoffmeister. 2022. *Wahrnehmungsdialektologie*. De Gruyter, Berlin, Boston.

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.

W. F. Sendlmeier and J. Seebode. 2006. Formantkarten des deutschen Vokalsystems. https://www.kw.tu-berlin.de/fileadmin/a01311100/Formantkarten_des_deutschen_Vokalsystems_01.pdf.

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.

Matthias J Sjerps, Neal P Fox, Keith Johnson, and Edward F Chang. 2019. Speaker-normalized sound representations in the human auditory cortex. *Nature communications*, 10(1):2465.

Georg Wenker. 2013. *Schriften zum "Sprachatlas des Deutschen Reichs". Gesamtausgabe*. Number 111.1–3 in Deutsche Dialektgeographie. Olms, Hildesheim, New York, Zürich.

9. Language Resource References

Paul Boersma and David Weenink. 2022. *Praat: doing phonetics by computer [Computer program]*. Version 6.2.14, retrieved 24 May 2022.

Ramon Corretge. 2012–2024. *Praat Vocal Toolkit [Software extension for Praat]*. Retrieved 20 January 2024.

Yannick Jadoul and Bill Thompson and Bart de Boer. 2018. *Parselmouth: A Python interface to Praat [Software]*. Version 0.4.1, retrieved September 2022.

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