Does Preprocessing Matter? An Analysis of Acoustic Feature Importance in Deep Learning for Dialect Classification

Lea Fischbach[1] and **Caroline Kleen**[1] and **Lucie Flek**[2,3] and **Alfred Lameli**[1]
[1] Research Center Deutscher Sprachatlas, Philipps-Universität Marburg
{lea.fischbach, caroline.kleen, lameli}@uni-marburg.de
[2] Bonn-Aachen International Center for Information Technology (b-it), University of Bonn
[3] The Lamarr Institute for Machine Learning and Artificial Intelligence
flek@bit.uni-bonn.de

Abstract

This paper examines the effect of preprocessing techniques on spoken dialect classification using raw audio data. We focus on modifying Root Mean Square (RMS) amplitude, DC-offset, articulation rate (AR), pitch, and Harmonics-to-Noise Ratio (HNR) to assess their impact on model performance. Our analysis determines whether these features are important, irrelevant, or misleading for the classification task. To evaluate these effects, we use a pipeline that tests the significance of each acoustic feature through distortion and normalization techniques.

While preprocessing did not directly improve classification accuracy, our findings reveal three key insights: deep learning models for dialect classification are generally robust to variations in the tested audio features, suggesting that normalization may not be necessary. We identify articulation rate as a critical factor, directly affecting the amount of information in audio chunks. Additionally, we demonstrate that intonation, specifically the pitch range, plays a vital role in dialect recognition.

1 Introduction

In the realm of deep learning, preprocessing plays a crucial role in optimizing model performance. While many studies focus on text, like (Uysal and Gunal, 2014), others concentrate on Environmental Sound Classification (ESC) or Automatic Speech Recognition (ASR) (Pfau et al., 2000). For instance, Bansal and Garg (2022) are exploring existing papers on preprocessing for ESC. Additionally, some studies focus on using spectrograms for audio processing (Chaiyot et al., 2021). Moreover, some research has attempted to enhance speech recordings for dialect identification, leading to improved subjective quality (Kakouros et al., 2020). However, these studies did not evaluate whether such preprocessing techniques actually improve the performance in downstream tasks.

Furthermore, despite studies such as (Lounnas et al., 2022), which incorporate noise reduction as a preprocessing step for dialect identification, a comprehensive study on the key aspects of audio preprocessing for dialect identification remains lacking. Often, only individual aspects of preprocessing are considered, as seen in (Pfau et al., 2000), where vocal tract length normalization (VTLN) and speech rate normalization (SRN) are examined.

Large-scale systems such as Whisper (Radford et al., 2023) and Meta's Massively Multilingual Speech (MMS) project (Pratap et al., 2024) highlight the power of extensive and diverse datasets in advancing ASR and language identification. Whisper, trained on 680,000 hours of multilingual and multitask supervised data, achieves improved robustness to accents, background noise, and technical language, demonstrating the impact of its large dataset. Similarly, Meta's MMS project tackles the lack of ASR systems for many languages by using religious texts, translated into numerous languages, to build a diverse training dataset. These projects showcase the importance of large datasets for robustness and inclusivity. In contrast, this study addresses the challenges of working with smaller, constrained datasets.

Notably, no paper has been found that investigates the effects of preprocessing raw audio on language or dialect classification. This gap is particularly significant in the context of deep learning-based dialect identification (DID), where understanding the fundamental aspects of audio preprocessing tailored specifically for dialect classification remains under-explored. This issue resonates with

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 159–169
March 3-4, 2025 ©2025 University of Tartu Library

findings in music information retrieval research, where deep learning efforts often prioritize optimizing hyperparameters that define network structure, while the audio preprocessing stage is often not optimized (Choi et al., 2018).

This study aims to bridge this gap by investigating how preprocessing adjustments affect the performance of dialect classification models trained on German audio data. We concentrate on the raw waveform and underscore the importance of different audio features in dialect classification. Specifically, we aim to determine whether adapting audio inputs improves model performance and whether certain features are misleading for the model, causing it to learn irrelevant patterns. Additionally, we explore if deep learning models inherently learn to ignore such variations or if performance even worsens, indicating that these features are important for dialect recognition in German.

Our contributions are threefold:

- We demonstrate that deep learning models for dialect classification are immune to variations in the tested audio features, suggesting that normalizations are not necessary.
- We reveal that the amount of information in an audio chunk is related to the Articulation Rate, impacting model performance.
- We show that intonation, specifically the pitch range within an audio chunk, is important for dialect recognition.

To achieve these contributions, we employ a pipeline that analyzes the significance of various acoustic features, representing a novel approach in the field.

By focusing on these aspects, our work not only fills a significant gap in the existing literature but also provides valuable insights for future research and applications in dialect classification using deep learning.

2 Used Acoustic Features

Used Acoustic Features are **Root Mean Square (RMS) amplitude**, **DC-offset**, **Articulation Rate (AR)**, **Pitch**, and **Harmonics-to-Noise Ratio (HNR)**.

RMS amplitude of a digital audio signal represents its perceived loudness and is simultaneously the mean absolute value of the signal. While RMS measures the average power of a signal, intensity in decibels (dB) quantifies the power relative to a reference level, typically the threshold of human hearing, on a logarithmic scale. As these metrics are correlated, only RMS is considered in this study.

RMS amplitude reflects both the speaker's vocal effort and external factors such as the recording equipment and the recording environment, including background noise and microphone distance.

DC-offset (also known as DC-bias), determined by the average amplitude of a segment of the signal, indicates a deviation from the symmetrical nature of a normal voice signal. In a typical symmetric sine signal, the high peak equals the low peak, resulting in an average value near zero over time. However, when a DC offset is present, the symmetry is disrupted, and the average value deviates from zero[1]. Despite being imperceptible, it reduces the available dynamic range, limiting the signals amplitude variation. DC-offset is primarily influenced by the recording equipment rather than the speaker.

Articulation Rate (AR) measures syllables per second during speech, excluding pauses, whereas Speech Rate (SR) includes pauses in its calculation. In this study, AR is emphasized over SR, as the audio data has been preprocessed to exclude pauses and non-articulatory elements. Also Otto (2012) states that variations in articulation speed between speakers may be more indicative of individual speaking styles than differences in overall speech tempo. The AR regarding to regional distribution has been minimally investigated thus far. Hahn and Siebenhaar (2016) found that there are differences in AR, but also suggest that this may correlate with other processes such as the elision of segments. They conclude that there must be different sound duration ratios in the different regions.

Pitch, often referred to synonymously as F0, stands for the fundamental frequency of a sound wave. F0 refers to the physical oscillation, while pitch denotes the perceived tonal height of the sound. In tools such as Praat (Boersma and Weenink, 2021), the pitch refers to F0. Pitch normalization, akin to Vocal Tract Length Normalization (VTLN), aims to mitigate speaker-specific variations in speech signals attributed to differences in vocal tract lengths, which are influenced by physiological factors such as sex. In explor-

[1] `https://solicall.com/dc-offset-and-audio-filtering/`

ing the connection between pitch and dialects, it's noteworthy that the typical fundamental frequency doesn't always align directly with dialect variations. Instead, phenomena such as variations in voice quality due to dialectal influences can affect pitch.

The **Harmonics-to-Noise Ratio (HNR)** quantifies the relationship between periodic components and noise in a signal. It measures acoustic periodicity by comparing the energy of harmonics to that of noise, with the result expressed in decibels (dB), indicating the dominance of periodic components over noise: an HNR of 20 dB signifies 99% of energy in periodic components and 1% in noise, calculated as $10 * log10(99/1)$. An HNR of 0 dB indicates equal energy distribution between harmonics and noise[2]. In speech analysis, HNR is favored over Signal-to-noise ratio (SNR) for its ability to capture voiced sounds periodicity. HNR primarily reflects characteristics of the speaker's voice, such as vocal cord vibration regularity and voice quality, but can also be affected by the recording equipment and environmental noise.

3 Experimental Setup

3.1 Used Corpus

This study utilizes automatically segmented audio files (Fischbach, 2024) sourced from the "Regionalsprache.de" (REDE) corpus (Schmidt et al., 2020ff.). The REDE corpus, which consists exclusively of recordings from male speakers, includes recordings from three age groups: young (18-23 years), middle-aged (45–55 years), and older (65+ years) speakers, captured across five different recording situations[3].

However, for the purposes of this study, only data from the older generation (65+ years) is analyzed. They are chosen due to their presumed higher dialect competence and to save computing time using only one generation. Furthermore, we only utilize the so-called dialectal "Wenker Sentences"[4] from the corpus. In this recording situation, an interviewer reads 40 sentences in Standard German, and the dialectal speakers translate these sentences into their local dialect. In total there are around 18 hours of audio data from the older generation and this recording situation, consisting of audios featuring only the dialectal speakers.

For classification we analyze a total of 20 different German dialects, classified according to Wiesinger (1983) without the transition areas between dialects. Dialects with insufficient variance (less than 3 speakers per dialect) are not further considered.

3.2 Classification Pipeline

The described pipeline is available and visualized on GitHub[5]. Initially, all audio files are preprocessed to standardize their format by converting them to mono, adjusting the bit-depth to 16 bits, and setting the sampling rate to 16 kHz, in line with the specifications of Google's TRILLsson models (Shor and Venugopalan, 2022), which is used for embedding extraction. The audio files are then divided into 10-second chunks for the extraction of these embeddings. Prior tests have shown this duration to be optimal. Shorter chunks yielded significantly poorer results, likely due to insufficient contextual information, whereas longer chunks offered no further gains, as the additional information in extended audio segments made 10 seconds sufficient. The resulting embeddings are processed through a small convolutional neural network (CNN) consisting of three dense layers with LeakyReLU activations and dropout layers to prevent overfitting. The network is trained using the Adam optimizer (Kingma and Ba, 2015). For model validation and testing, $\lceil \frac{\#S_D}{10} \rceil$ speakers are randomly selected from each dialect, where $\#S_D$ represents the total number of speakers in the respective dialect. To account for variability in results due to different speaker selections, we employ a Monte Carlo cross-validation approach, repeating the data splitting and model evaluation process 250 times with new random speaker selections in each run. This number of iterations was chosen based on prior tests demonstrating its effectiveness in detecting significant differences between experiments. The mean of the weighted F1-score across runs is calculated, and the Mann-Whitney U test (Mann and Whitney, 1947) is used to assess the statistical significance of performance differences between runs.

[2] https://www.fon.hum.uva.nl/praat/manual/Harmonicity.html

[3] Additional information about the recording situations, the recording locations and the project itself can be found on https://rede-infothek.dsa.info/

[4] https://www.uni-marburg.de/en/fb09/dsa/research-documentation-center/wenkersaetze

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

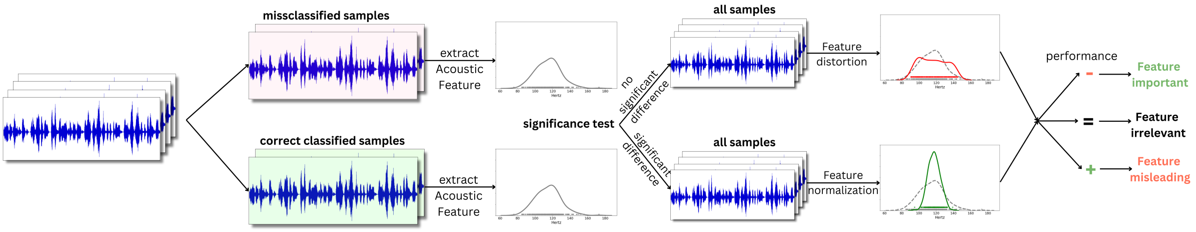


Figure 1: Visualization of the procedure.

3.3 Procedure

The entire procedure is summarized in the pipeline diagram shown in Figure 1, which outlines the steps involved in evaluating the significance of different features for dialect classification. Initially, we conduct analyses to determine significant differences in the distribution of feature values between two groups of audio chunks: those with a high misclassification rate (misclassified in 95%-100% of 250 runs, referred to as ”wrongly classified”) and those that are frequently classified correctly (correctly classified in 65%-100% of 250 runs, referred to as ”correctly classified”). These thresholds are chosen to ensure a balanced representation of both incorrect and correct chunks. This can be inferred from the diagram in Figure 2. The diagram illustrates how many chunks are classified correctly and incorrectly at which percentage threshold, and the ratio of the number of incorrect to correct chunks.

Model performance is evaluated using the weighted F1-score to account for the imbalanced class distribution. If a significant difference is found between the distributions of features (such as pitch) for incorrectly and correctly classified chunks — determined using the Mann-Whitney U test, where a p-value < 0.05 indicates a significant difference — all chunks are normalized according to that specific feature. If the difference is not significant, the chunks are deliberately distorted to assess whether this manipulation affects the model’s performance. This approach helps to identify whether a particular feature is important, irrelevant, or even misleading for the deep learning model in classifying (German) dialects. The rationale behind these assessments, such as why feature distortion resulting in decreased model performance indicates the feature’s importance, is summarized in Table 1.

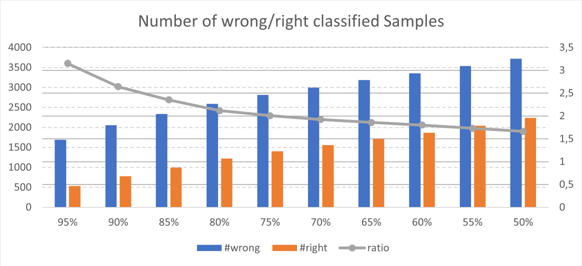


Figure 2: Ratio and number of wrongly and correctly classified chunks with percentage threshold values.

3.4 Feature Extraction

The features are computed using Praat (Boersma and Weenink, 2021) via the Parselmouth Python interface (Jadoul et al., 2018), which facilitates Praat script execution in Python, as detailed below:

RMS Amplitude Extraction: The RMS value is computed using Praat’s `Get root-mean-square` function.

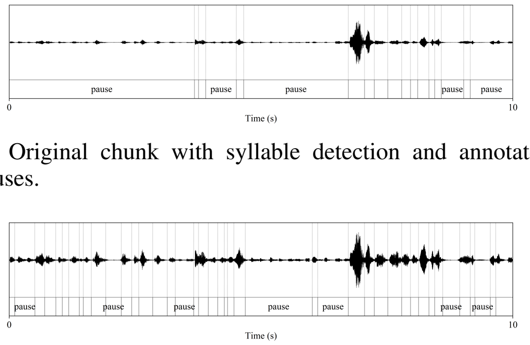
DC-Offset Extraction: The DC-offset is calculated as the mean value of the audio chunk.

AR Extraction: We employ a multi-step process to extract the articulation rate. Initially, all audio chunks are peak-normalized to standardize intensity levels, enhancing the accuracy of the syllable recognition algorithm. Following normalization, we use a Praat script from the *Praat Vocal Toolkit* (Corretge, 2012-2024), which identifies syllable nuclei while discarding non-voiced peaks, to mark syllables in the audio segments. This Praat script is described by De Jong and Wempe (2009). The articulation rate is then extracted as the ratio of the number of syllables to the phonation time.

Peak normalization before AR extraction is needed, to address the incorrect detection of pauses in audio chunks where none should exist. As illustrated in Figure 3 a), many pauses were falsely identified in places where speech is present due to fluctuations in intensity, highlighting the

Feature Distortion			Feature Normalization		
-	=	+	-	=	+
Distortion degrades performance as the model relies on the original distribution.	The Feature is irrelevant; distortion has no effect.	Distortion removes misleading information, improving performance.	Normalization removes useful distinctions, degrading performance.	Differences in distribution are irrelevant; no effect.	Normalization reduces noise or bias, improving performance.

Table 1: Effects of Feature Distortion and Normalization on model performance (+ improved, - degraded, = unchanged), indicating the role of the Feature in the model, as in Figure 1.



(a) Original chunk with syllable detection and annotated pauses.

(b) Same chunk after peak normalization with syllable detection and annotated pauses.

Figure 3: Audio chunk and its detected syllables/pauses.

algorithm's sensitivity to sound levels. Peak normalization improves syllable recognition by stabilizing these fluctuations, which is particularly beneficial for speech rate (SR) detection, as it is more affected by misclassified pauses compared to AR, but it also improves AR performance.

Figure 3 b) shows the same audio chunk after peak normalization was applied before running the syllable recognition algorithm. Although some incorrect pauses remain, the results are significantly more accurate. Across all chunks, this process reduced the standard deviation of the ratio of speaking duration to audio length (which ideally should be close to 1 for our dataset, as there should be no or only very short pauses), bringing more values closer to 1 and minimizing extreme deviations.

Additionally, our use of 10-second chunks ensures a stable extraction of AR, aligning with findings from Arantes and Lima (2017), where they state that both SR and AR stabilize after approximately 9 seconds.

Pitch Mean and Pitch Standard Deviation Extraction: We calculate both the mean pitch and the standard deviation of the pitch restricting the analysis to a range of 80 Hz to 170 Hz. Pitch values are extracted using the *To Pitch* function in Praat, followed by the computation of either the mean or the standard deviation.

The pitch range of 80-170 Hz is selected, because Praat's default settings for pitch extraction often result in high pitches values (up to 240 Hz) and large fluctuations (up to 100 Hz within a chunk), leading to a high standard deviation (±44.74 Hz). This issue is likely due to flaws in the underlying algorithm (Boersma et al., 1993). Adjusting the pitch range to match the typical frequency range for the speaker's sex (and age) can mitigate this problem and ensures more accurate pitch detection.

The default pitch range in Praat is set between 75 Hz and 600 Hz. This range can be narrowed to 80-170 Hz, which corresponds to the normal pitch range for male speakers. For instance, a study involving 2472 German-speaking men aged 40–79 years found that the mean fundamental frequency of the conversational speaking voice was 111.9 Hz, with specific averages of 112.9 Hz (±17.5) for ages 60–69 and 120.6 Hz (±19.8) for ages 70–79 (Berg et al., 2017). Another study reported a mean pitch of 120 Hz (±18 Hz) for the German male reading voice (Andreeva et al., 2014). Our adjusted pitch range of 80–170 Hz is therefore well-supported by these findings.

With the new settings, the largest deviation within a chunk decreased by nearly one-third to 34.13 Hz, and the standard deviation was reduced by more than half to 19.09 Hz. Figure 4 visualizes the results of pitch extraction using the two

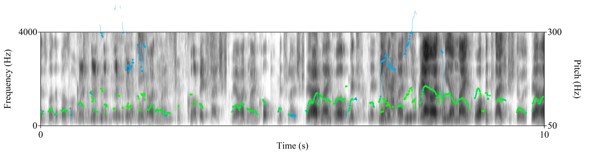


Figure 4: Extracted Pitch with Praat for different pitch ceilings: 75-600 Hz (blue) and 80-170 Hz (green).

ranges: the extracted pitch using standard settings (75-600 Hz) is shown in blue and in green with adjusted settings (80-170 Hz). It is evident that the blue pitch values are often too high, while the green ones are much more stable. By narrowing the pitch range, the algorithm is constrained to estimate the pitch within these plausible bounds, thereby providing results closer to the true pitch.

HNR Extraction: The HNR gets extracted using Praat's `To Harmonicity (cc)` function.

4 Results

Table 2 summarizes the feature importance analysis. The *p-val* column displays the p-values from Mann-Whitney U tests, comparing feature values between correctly and wrongly classified chunks; a p-value below 0.05 indicates a statistically significant difference. The *Norm./Dist.* column indicates whether the audio chunks were normalized or disturbed, with the *Method* column detailing the specific processing method. The *new Perf.* column presents the mean weighted F1-Score of the model with altered audio chunks, compared to the original score of 0.228. The *Perf. p-val* column contains the p-value from the Mann-Whitney U test comparing the model's performance with original versus altered chunks, indicating the feature's impact on performance, as shown in the *Feat. Imp.* column.

4.1 RMS

In the analysis of Root Mean Square (RMS) amplitude, statistically significant distinct distributions can be observed between wrongly and correctly classified chunks with a lower mean RMS for wrongly classified chunks. So RMS gets normalized for each chunk to assess the impact of RMS values on classification results. Peak normalization, which adjusts audio signals relative to their loudest point and has been set so that the highest peak reaches -0.2, and loudness normalization, which aims to standardize perceived loudness, are explored. Where for loudness normalization there is a risk of clipping if the target value is set to high. Both normalization methods resulted in an overall increase in loudness, as depicted in Figure 5 a), with loudness normalization demonstrating a notably reduced standard deviation due to its uniform adjustment to a target loudness level of -14dB.

Neither of the two methods yields a significant difference in classification performance. In both normalization methods, approximately the same errors are observed in assigning chunks to dialects as without normalization.

This finding supports the theory that the initially different distribution of correctly and incorrectly classified chunks was coincidental. Further tests have shown that speakers with almost the same misclassification rate from the same dialect can have very different RMS, likely due to varying recording conditions and consequently different RMS levels. Therefore, the differences in the distributions of correctly and incorrectly classified chunks are likely due to variations in the classification performance of individual speakers, rather than differences in RMS levels. Therefore, it is reasonable to conclude that the volume of individual chunks does not influence the model's performance, rendering this feature irrelevant.

4.2 DC-Offset

There is a significant difference in distributions of wrongly and correctly classified chunks. To address this, a normalization technique called *Mean Centering* is employed by subtracting the mean of each chunk, effectively minimizing the DC-Offset. However, this normalization yields no difference in classification performance.

Yet, since even the largest offset in our data is minimal (-0.0007), it should be tested again whether a disturbance of the DC-Offset leads to a deterioration in classification performance. Each chunk gets a randomized disturbance within the range of [-0.1, 0.1], which also can be seen in Figure 5 b) at the bottom. However, this perturbation fails to yield any discernible difference in model performance, suggesting that the DC-offset holds little relevance to classification performance, as long as it is in normal ranges.

Feature	p-val	Norm./Dist.	Method	new Perf.	Perf. p-val	Feat. Imp.
RMS	0.000	Norm.	RMS von -14dB	0.287	0.104	=
			Peak of -0.2	0.283	0.306	=
DC-Offset	0.000	Norm.	Mean Centering	0.280	0.635	=
		Dist.	[-0.1, 0.1]	0.283	0.321	=
AR	0.936	Dist.	[2.5, 6] Syll./Sec.	0.259	0.000	-
		Norm.	4.3 Syll./Sec.	0.277	0.870	=
Pitch mean	0.235	Dist.	[90, 160] Hz	0.276	0.626	=
Pitch std	0.012	Norm.	Half orig. Std.	0.236	0.001	-
			Monotonized	0.241	0.000	-
			Std. of 18	0.269	0.087	=
HNR	0.000	Norm.	Praat	0.283	0.127	=
			Noisereduce	0.270	0.419	=

Table 2: Summary of feature importance analysis, including p-values from Mann-Whitney U tests comparing feature values between correctly and incorrectly classified chunks (p-val), processing methods applied to normalize or disturb features (Norm./Dist. and Method), mean weighted F1-Score with altered audio chunks, p-values comparing model performance with original and altered audio chunks (Perf. p-val), and the resulting feature importance (Feat. Imp.).

4.3 AR

The distributions of the wrongly and correctly classified chunks are similar, so AR perturbation is conducted to assess its impact on model performance. AR values are intentionally disturbed between 2.5 and 6 syllables per second, derived from extreme measurements from all chunks as can be seen in the top box plot of Figure 5 c). Model performance significantly declines with disturbed AR chunks compared to normal conditions. To ascertain whether this decline stems solely from extreme differences between chunks or generally from extreme AR values, additional tests are conducted. These involve assessing the model's behavior with only slow (AR of 3.0) or fast (AR of 6.0) chunks. When the articulation rate is reduced, resulting in slower audio, the chunks are still formed with a fixed length of 10 seconds. As a result, there are more chunks overall, but each chunk contains less information due to the lower tempo. With higher AR, there are fewer chunks, but each chunk contains more information. Chunks with lower AR lead to a 44% increase in length and degraded model performance. Conversely, chunks with higher AR, approximately 71.36% shorter than the original recordings, are showing no significant difference. Considering that longer chunks generally contain more information, this could explain why the classification performance did not deteriorate or improve with a faster AR, as the model's performance in earlier tests also did not benefit from chunks longer than 10 seconds. Moreover, reducing the length of chunks with the higher AR up to 8 seconds does not yield a significant difference in model performance. However, caution should be exercised not to increase the AR too much. Another test using chunks of 7 seconds with double the AR compared to the original mean, resulting in an AR of 8.734, shows a significant deterioration. These findings suggest that, to a certain extent, manipulating AR to increase chunk speed can be effective in shortening chunk length for reduced computational workload without compromising classification performance. This approach has the potential to conserve computational resources during classification tasks. Nevertheless, the tradeoff between increased AR and reduced audio length is limited. If the speech speed is too slow, longer chunks should instead be used to ensure sufficient information is captured. Therefore, it is assumed that AR does not influence dialect classification, but rather the amount of information contained in each chunk. This also agrees with De Jong and Wempe (2009) where they state that speech recognizers perform relatively poorly when speech rate is very fast or very slow.

Nevertheless, we aim to normalize the AR, as suggested by Pfau et al. (2000), where Speech Rate Normalization resulted in a reduction of word error rate. To normalize the AR, all chunks are speed-manipulated based on their original AR. The median AR across all audio chunks is 4.36 and the mean is 4.37. Therefore, all audios should have an articulation rate of approximately 4.3. To achieve that, first, the factor between the current and the desired AR is determined using $factor = AR_{old}/AR_{new}$, and then the original audio chunk is speed-manipulated by this factor (factor < 1 results in the audio being faster). Through this approach, the AR is slightly reduced on average, resulting in a slowdown of most audios, as can be seen at the lower mean in Figure 5 c). Normalizing the AR had no impact on the classification performance.

4.4 Pitch

Since there is no significant difference between wrongly and correctly classified chunks where values are extracted with the adjusted pitch range, further testing is conducted to determine if the model's performance would degrade when the pitch is randomly altered. The pitch is varied between 90-160 Hz, a range considered normal for male speaking voice and providing headroom in both directions for pitch extraction with Praat. Despite this manipulation, no significant difference in classification performance can be observed. These findings indicate that pitch does not significantly impact classification.

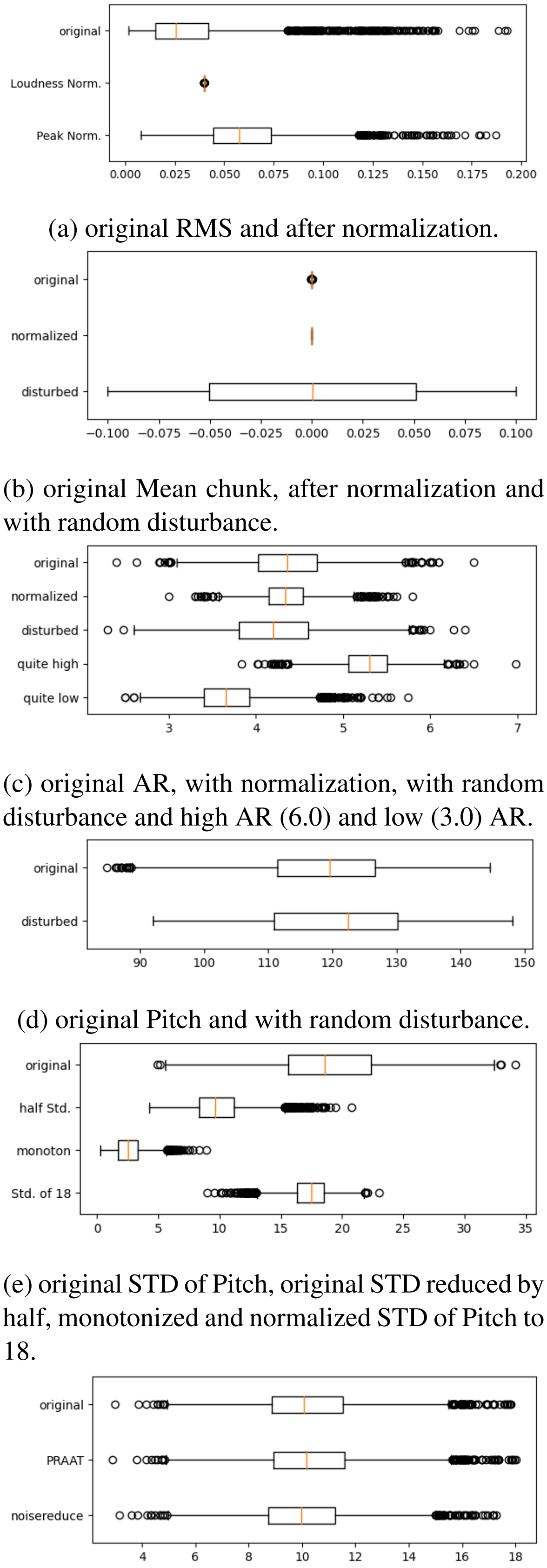
Additionally, we investigate how the model behaves when adjusting the pitch range by altering the standard deviation. The distribution of pitch standard deviations between correctly and incorrectly classified chunks differs significantly. Specifically, the mean standard deviation of pitch is higher for incorrectly classified chunks. To address these differences, we normalize the pitch range of each chunk. We test several approaches: halving the pitch range, monotonizing the pitch, and normalizing it to a standard deviation of 18 Hz. The model's performance significantly deteriorate when the pitch range is halved or monotonized, while normalization to 18 Hz standard deviation shows no impact on performance. These results, depicted in Figure 5 e), suggest that pitch variation is important for dialect classification, but only its intonation (the variance in pitch) rather than its exact magnitude.

The importance of pitch in language and dialect classification is further highlighted by Vicenik and Sundara (2013), where features such as minimum, maximum, and mean pitch, as well as the number and characteristics of pitch rises, were used to distinguish between German and American English with an accuracy of 86%. Notably, these features primarily captured pitch variance rather than absolute values, emphasizing the significance of intonation. The study also demonstrated that these pitch features could differentiate between varieties of English, such as American and Australian English, achieving an accuracy of 79%. Their study also concludes that intonation plays a crucial role in helping listeners distinguish between different languages.

4.5 HNR

Statistically significant differences in the distributions of HNR values are observed between correctly and wrongly classified chunks, with the mean HNR slightly higher for the latter. As illustrated in Figure 5 f), the majority of our audio chunks have a mean HNR value indicating approximately 90% harmonic content, though some chunks exhibit lower HNR values with around 70% harmonic content. Due to the absence of stationary background noises applying a constant bandpass filter is not feasible. Furthermore, since the recordings were downsampled to 16 kHz, any noise above 8 kHz is already filtered out.

Attempts to reduce noise using Praat's spectral subtraction method, as defined by Boll (1979), yields no significant changes in HNR or improvements in classification performance. We also applied the `noisereduce` library (Sainburg et al., 2020; Sainburg, 2019). Non-stationary noise reduction is applied due to the absence of specific interfering noises, yet this also results in minimal changes in HNR and no significant difference in classification. This result is consistent with findings from Lounnas et al. (2022), where noise reduction using `noisereduce` showed no notable effect on classification performance when using Convolutional Neural Networks (CNNs). Thus, it can be concluded that non-stationary noises have little to no influence on the performance of dialect classification, as long as they do not obscure the speech signal.



(a) original RMS and after normalization.

(b) original Mean chunk, after normalization and with random disturbance.

(c) original AR, with normalization, with random disturbance and high AR (6.0) and low (3.0) AR.

(d) original Pitch and with random disturbance.

(e) original STD of Pitch, original STD reduced by half, monotonized and normalized STD of Pitch to 18.

(f) original HNR, HNR reduced with Praat and HNR reduced with noisereduce.

Figure 5: Boxplots for different Audio-chunk features.

5 Discussion

The study's results indicate that pitch variation did not impact model performance among the used group of older males, suggesting its potential applicability across different age groups. However, it's uncertain if pitch normalization would have the same effect in a diverse group, where sex and age may introduce more pitch variation. Future studies should explore the impact of pitch normalization on mixed demographics and evaluate broader techniques such as voice conversion techniques to standardize all audio inputs.

Although the analysis focused on a German dialect dataset, these insights could extend to other corpora. Nonetheless, it is essential to conduct a thorough evaluation of each dataset's features to ensure that preprocessing techniques are well-suited to its specific characteristics and contribute to the classification tasks coherence and relevance.

The precise feature extraction values may vary depending on the extraction methods and parameters used. However, RMS and DC-offset measurements should consistently yield the same results, as these values can be accurately calculated. In contrast, when extracting pitch features, parameters such as the pitch floor and ceiling should be adjusted according to the age and sex of the speakers to obtain more accurate estimations of the true pitch.

Acknowledgements

This research is supported by the Academy of Science and Literature Mainz (grant REDE 0404), the German Federal Ministry of Education and Research (BMBF) (grant AnDy 16DKWN007), 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 thank the anonymous reviewers for their valuable feedback.

References

Bistra Andreeva, Grazyna Demenko, Magdalena Wolska, Bernd Möbius, Frank Zimmerer, Jeanin Jügler, Magdalena Oleskowicz-Popiel, and Jürgen Trouvain. 2014. Comparison of pitch range and pitch variation in Slavic and Germanic languages. In *Proceedings to the 7th Speech Prosody Conference. Trinity College Dublin, Ireland. May 20-23, 2014*, pages 776–780. International Speech Communication Association.

Pablo Arantes and Verônica Gomes Lima. 2017. Towards a methodology to estimate minimum sample length for speaking rate. *Revista do GEL*, 14(2):183–197.

Anam Bansal and Naresh Kumar Garg. 2022. Environmental Sound Classification: A descriptive review of the literature. *Intelligent Systems with Applications*, 16:200115.

Martin Berg, Michael Fuchs, Kerstin Wirkner, Markus Loeffler, Christoph Engel, and Thomas Berger. 2017. The Speaking Voice in the General Population: Normative Data and Associations to Sociodemographic and Lifestyle Factors. *Journal of Voice*, 31(2):257.e13–257.e24.

Paul Boersma and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 `http://www.praat.org/`.

Paul Boersma et al. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Amsterdam.

Steven Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120.

Krittaya Chaiyot, Supattra Plermkamon, and Thana Radpukdee. 2021. Effect of audio pre-processing technique for neural network on lung sound classification. In *IOP Conference Series: Materials Science and Engineering*, volume 1137. IOP Publishing.

Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. 2018. A comparison of audio signal preprocessing methods for deep neural networks on music tagging. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1870–1874. IEEE.

Ramon Corretge. 2012-2024. Praat Vocal Toolkit. retrieved 20 January 2024 `https://www.praatvocaltoolkit.com`.

Nivja H De Jong and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390.

Lea Fischbach. 2024. A Comparative Analysis of Speaker Diarization Models: Creating a Dataset for German Dialectal Speech. In *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*, pages 43–51.

Matthias Hahn and Beat Siebenhaar. 2016. Sprechtempo und Reduktion im Deutschen (SpuRD). *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2016*, pages 198–205.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Sofoklis Kakouros, Katri Hiovain, Martti Vainio, and Juraj Šimko. 2020. Dialect identification of spoken North Sámi language varieties using prosodic features. In *Speech Prosody 2020*, pages 625–629.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Khaled Lounnas, Mohamed Lichouri, Mourad Abbas, Thissas Chahboub, and Samir Salmi. 2022. Towards an Automatic Dialect Identification System using Algerian Youtube Videos. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 258–264.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Christina Otto. 2012. Sprechgeschwindigkeit und Geschlechterunterschiede. *Phonetik & Phonologie 8 Friedrich-Schiller-Universität Jena*, 92(1):33.

Thilo Pfau, Robert Faltlhauser, and Günther Ruske. 2000. A combination of speaker normalization and speech rate normalization for automatic speech recognition. *Proc. 6th International Conference on Spoken Language Processing (ICSLP 2000)*, 4:362–365.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Tim Sainburg. 2019. timsainb/noisereduce: v3.0.0. retrieved 12 February 2021.

Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10).

Jürgen Erich Schmidt, Joachim Herrgen, Roland Kehrein, and Alfred Lameli. 2020ff. Regionalsprache.de. Forschungsplattform zu den modernen Regionalsprachen des Deutschen. Forschungszentrum Deutscher Sprachatlas Marburg.

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

Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112.

Chad Vicenik and Megha Sundara. 2013. The role of intonation in language and dialect discrimination by adults. *Journal of Phonetics*, 41(5):297–306.

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.