Identifying the features predicting perceived prominence in German and Catalan

Aleksandra Ćwiek, report from 2024-10-07

1. The aim and the features

I aimed to **predict perceived prominence using various acoustic features**. All features were extracted for the pre-, post-, and tonic syllables.

The acoustic features and their explanation:

duration: Measures the length of time an interval lasts.

duration_noSilence: The length of interval excluding silent intervals.

ampl_median: Median amplitude (correlate of loudness) of the sound.

ampl_noSilence_median: Median amplitude without silent intervals.

env_slope: The slope of the amplitude envelope, indicating the rate of change in amplitude envelope (correlate of loudness).

pitch_median_norm: Median fundamental frequency, normalized by speaker and language.

pitch_sd_norm: The standard deviation of fundamental frequency, indicating the variability of *f*0 within the interval.

f0_slope_norm: The slope of the normalized *f*0, indicating the rate of change within the interval.

specCentroid_median: The spectral centroid is a measure of where the center of mass of the spectrum is located. It provides an indication of the brightness of a sound. A higher centroid generally corresponds to a brighter sound, while a lower centroid indicates a darker or more muffled sound. It is calculated as the weighted mean of the frequencies in a signal, weighted by their magnitudes.

entropy_median: Entropy measures the disorder or unpredictability of a sound signal. Lower entropy indicates more structured and predictable sounds, while higher entropy signifies more randomness or noise. It is often used to characterize how much variability exists in a sound's frequency content.

HNR_median: HNR quantifies the ratio between the harmonic (periodic) components and the noise (non-periodic) components of a sound. A higher HNR indicates a clearer, more harmonic sound, often associated with voiced speech or clear tonal sounds. Lower HNR suggests more noise, such as in breathy or hoarse voices.

amEnvDep_median: This feature represents the deviation in the amplitude envelope of a sound over time. The amplitude envelope describes how the amplitude of the sound signal evolves. The amplitude envelope deviation reflects how much the amplitude deviates from its mean over time, providing insights into dynamic changes in loudness.

fmDep_median: Frequency modulation deviation measures how much the frequency of the signal deviates over time from a baseline frequency. It captures variations in pitch and is often used to detect vibrato or pitch fluctuations in speech and music. Larger deviations indicate more pronounced frequency shifts.

2. The mechanics of the models

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm based on decision tree ensembles. Like Random Forests, it uses multiple trees to improve predictive performance. However, XGBoost improves upon Random Forests by sequentially building trees, with each tree correcting the errors of the previous ones. This iterative process, known as "boosting", results in highly accurate models. XGBoost also incorporates regularization, which helps to control overfitting by penalizing overly complex models. This makes it particularly effective when dealing with datasets with many predictor variables, as it can identify and rank the most important features while maintaining model generalization. Additionally, its flexibility in tuning parameters allows it to be customized for optimal performance in various predictive tasks.

Given the presence of missing values for some of the features in our dataset, I explored two methods: omitting incomplete cases and imputing missing data with MICE (Multiple Imputation by Chained Equations). Based on initial results with random forests, I chose MICE, as it maintained data integrity and model accuracy better. Just briefly: MICE imputation iteratively fills in missing values by generating plausible estimates from the observed data. This method helps retain the statistical relationships between variables, resulting in a more accurate and robust dataset for modeling.

I divided the dataset into five subsets for cross-validation to ensure comprehensive model evaluation. Using the XGBoost algorithm, **I trained five models, each of them comprising of 80% of the data, and tested it on the remaining 20%**.

3. Results

Below I will discuss the results for the languages.

3.1. German

The **kappa values for our models ranged from 0.7045 to 0.7579**, indicating moderate to substantial agreement (kappa is like inter-rater agreement, so 0.7 is generally considered substantial). The **accuracy ranged from 0.8419 to 0.8623**, with **p-values less than 0.005**. Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. High accuracy means the model correctly predicts the outcome most of the time. This basically means that the models are performing very well and the results are trustworthy.

I extracted and normalized feature importance for each model, then calculated the cumulative importance by averaging the scores across all five models. The top 10 most-predictive features were (most predictive in the top):

Feature	**Cumulative Importance**
f0_slope_norm	0.2484
env_slopePost	0.1983
duration	0.1900
ampl_noSilence_median	0.1801
entropy_median	0.1735
entropy_medianPost	0.1698
pitch_sd_norm	0.1698

ampl_noSilence_medianPost	0.1612
pitch_median_norm	0.1581
env_slope	0.1565

I color-coded the pre-, post-, and tonic syllables.

Post means a post-tonic syllable, so for example, the slope of the amplitude envelope, the median entropy, and the median amplitude (without the silence intervals) in the post-tonic syllables is highly predictive of the rating. Other reliable predictors relate to the tonic syllable and to a large degree overlap with previous literature, except for entropy (which is also one in the post-tonic). So, **we have some corroborating and some novel results**! ☺

3.2. Catalan

The **kappa values for our models ranged from 0.7047 to 0.7431**, indicating moderate to substantial agreement. The **accuracy ranged from 0.8641 to 0.8842**, with **p-values less than 0.001**. This means the models are performing very well and the results are trustworthy.

Just as for German, here are the top 10 most predictive features, according to their cumulative importance:

Feature	**Cumulative Importance**
duration	0.2729
f0_slope_normPost	0.1769
pitch_sd_norm	0.1727
ampl_noSilence_median	0.1654
specCentroid_medianPost	0.1595
specCentroid_median	0.1537
pitch_median_norm	0.1536
specCentroid_medianPre	0.1526
HNR_medianPost	0.1525
pitch_median_normPre	0.1524

We have some **similarities to German**, namely duration, *f*0 (range and median), and median amplitude (without the silence intervals) of the tonic syllable are informative for the perceived prominence categorization. There are **also differences**, however; most notably instead of entropy, the spectral centroid median is generally informative – throughout pre-, post-, and tonic syllable! Also, more features of the surrounding (pre- and posttonic), rather than tonic syllable, are informative than in German.

3.3. Summary of the results across German and Catalan

For tonic syllables, both languages rely heavily on f0 and duration as key predictors of prominence. These features highlight that both the range of fundamental frequency and its average within the prominent syllable (reflected by *pitch_sd_norm* and *pitch_median_norm* respectively), as well as the overall length of the syllable, are essential in signaling prominence. Additionally, the median amplitude (without silence intervals) is predictive of perceived prominence in both German and Catalan, underscoring the importance of loudness during the prominence syllable.

In German, compared to Catalan, the slope of *f*0 and the slope of the amplitude envelope within the prominent syllable are among the most predictive features of prominence. This may indicate that in German, the rate of change (as represented by the steepness or flatness of the slope) in prosodically relevant features plays a more significant role than in Catalan. Furthermore, entropy, representing the randomness or complexity in the acoustic signal, is crucial in German, especially in both the tonic and posttonic syllables.

In contrast, Catalan identifies the median spectral centroid as a key feature. This feature measures the spectral center and corresponds to the perceived brightness of the sound. Importantly, it is relevant across all syllables – tonic, pretonic, and posttonic.

Catalan generally shows a stronger reliance on pretonic syllables. While no pretonic feature in German is found to be informative, Catalan identifies both the spectral centroid and f0 median of the pretonic syllable as reliably informative for predicting perceived prominence. These findings suggest that in Catalan, the acoustic properties leading up to the stressed syllable – particularly how the sound is balanced and how pitch unfolds – are vital for cuing prominence.

Regarding posttonic syllables, both languages have features that predict perceived prominence ratings, though these features differ between them. In German, the key posttonic features are the slope of the amplitude envelope, entropy, and the median amplitude (excluding silence intervals). In Catalan, the predictive posttonic features include *f*0 slope, the spectral centroid, and the harmonics-to-noise ratio (HNR). This suggests that the languages use different strategies in the posttonic environment: German places more importance on loudness and entropy, while Catalan relies more on frequency-related features.

These results imply that while both languages cue prominence similarly, with tonic syllables being central, they differ in how they utilize pretonic and posttonic syllables. Catalan's greater focus on pretonic syllables suggests a more gradual buildup to the stressed syllable, while German's balanced use of tonic and posttonic cues indicates a more direct approach to signaling prominence. This highlights subtle prosodic differences in how speakers of these languages perceive and produce stress in spoken discourse.

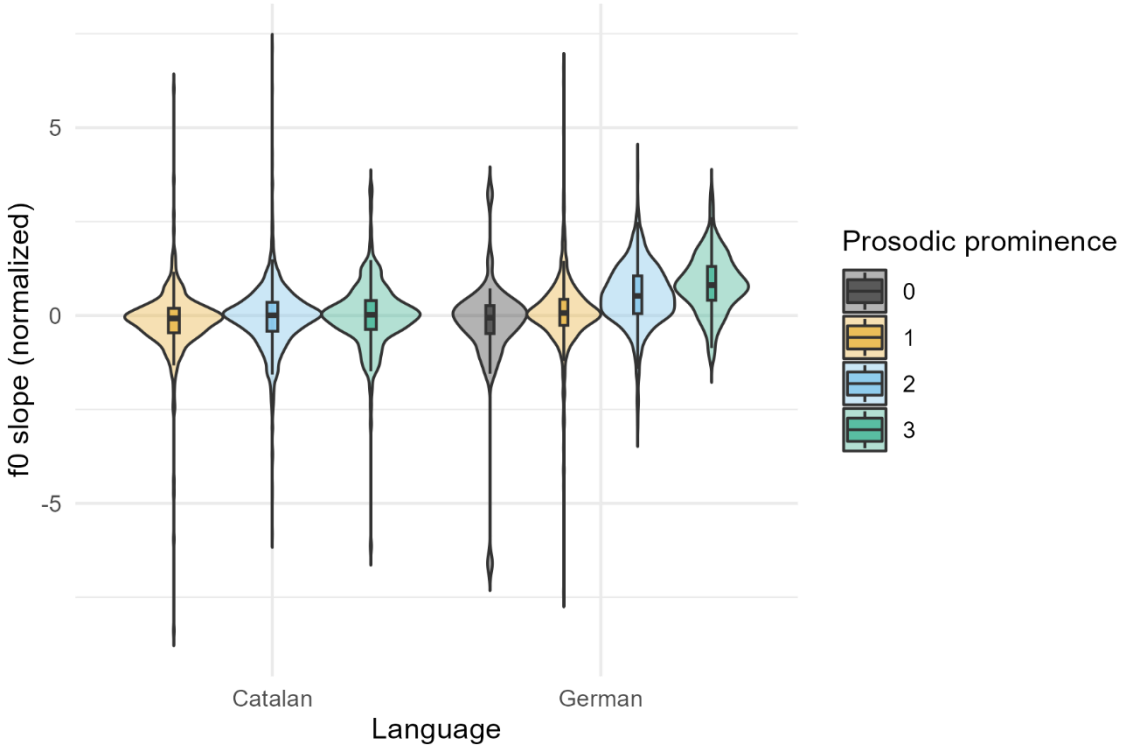
4. Plots

Below, I will present some of the plots that will also visualize the directionality of the effect. Please just focus on the given language, even though I always plot both. I present the plots in the order given by the cumulative feature importance, the value of which is given in the brackets next to the feature.

While you are looking at it, pay attention to how wide the values are spread in Catalan in comparison to German. This might suggest various things, for example individual differences across speakers or raters (also differences that raters have towards speakers, not towards features), or a mixture of these. The pattern in Catalan is definitely not as clear as in German.

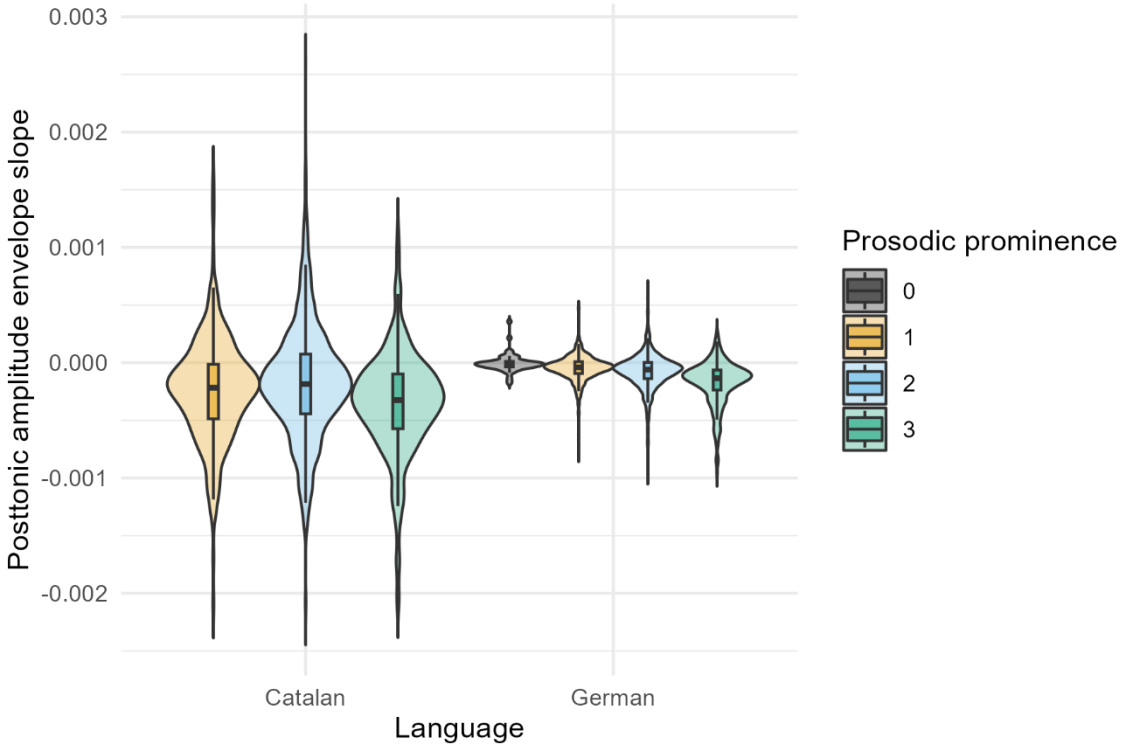
4.1. German

a. f0_slope_norm (0.2484)



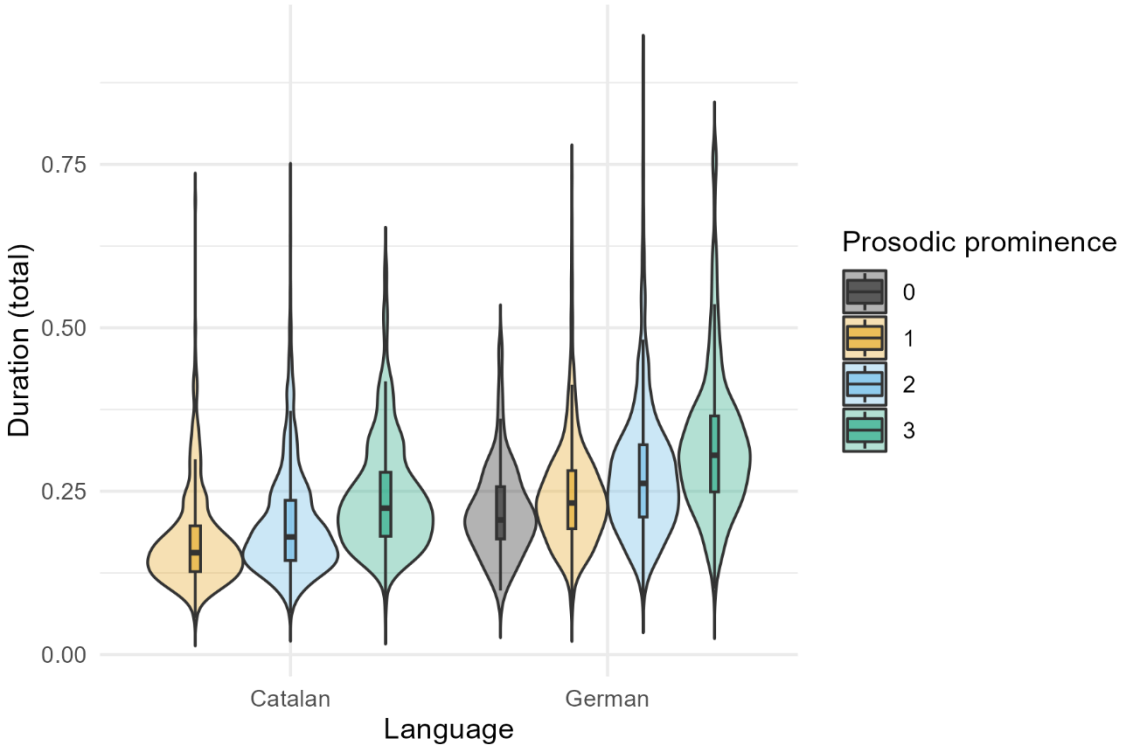
*F*0 slope in the tonic syllable is higher in higher perceived prosodic prominence.

b. env_slopePost (0.1983)



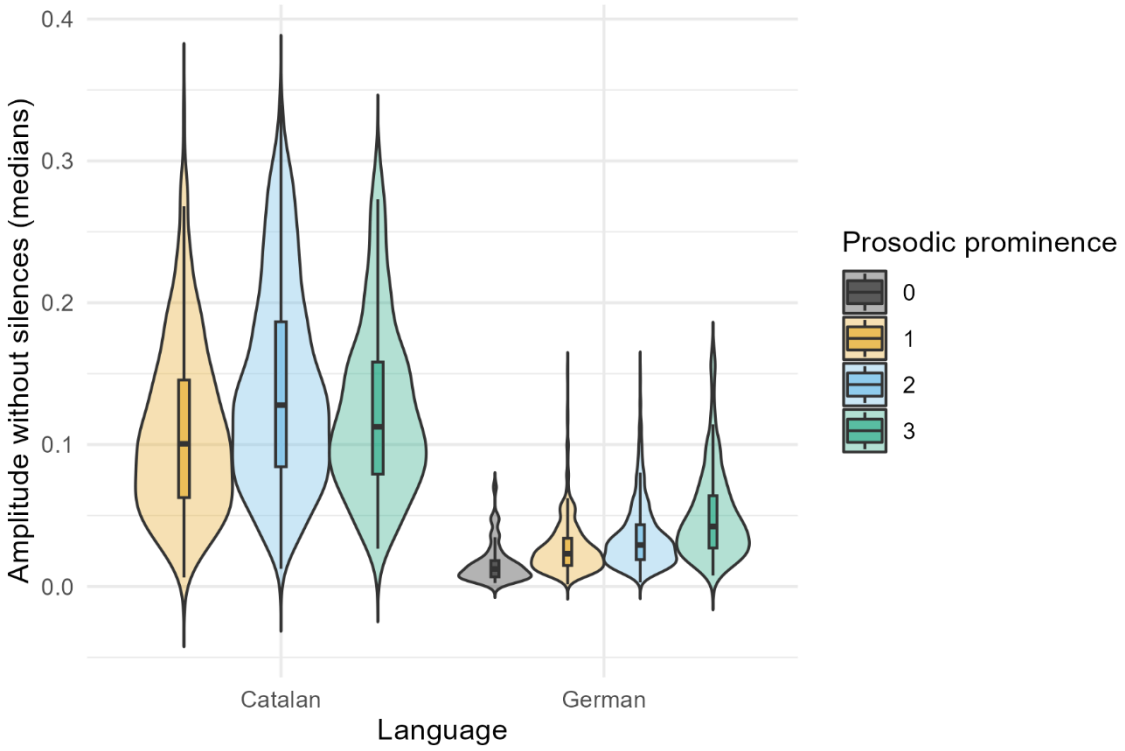
Amplitude envelope slope in the posttonic syllable is lower in higher perceived prosodic prominence.

c. duration (0.1900)



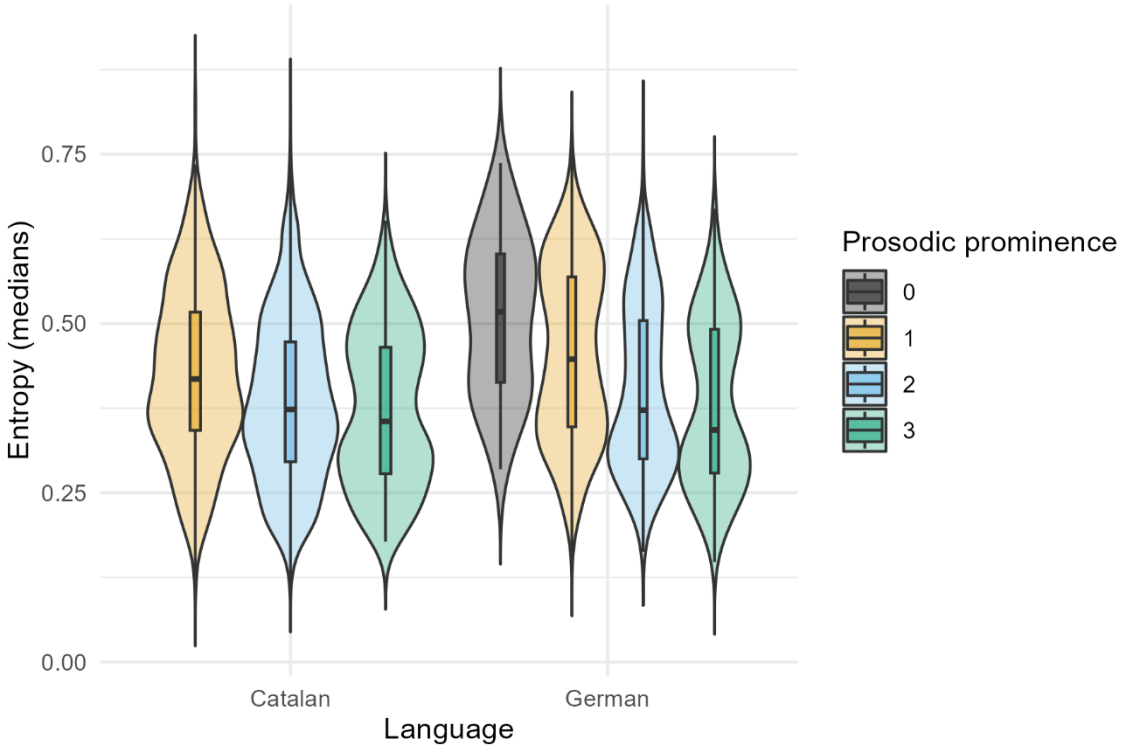
Duration in the tonic syllable is longer in higher perceived prosodic prominence.

d. ampl_noSilence_median (0.1801)



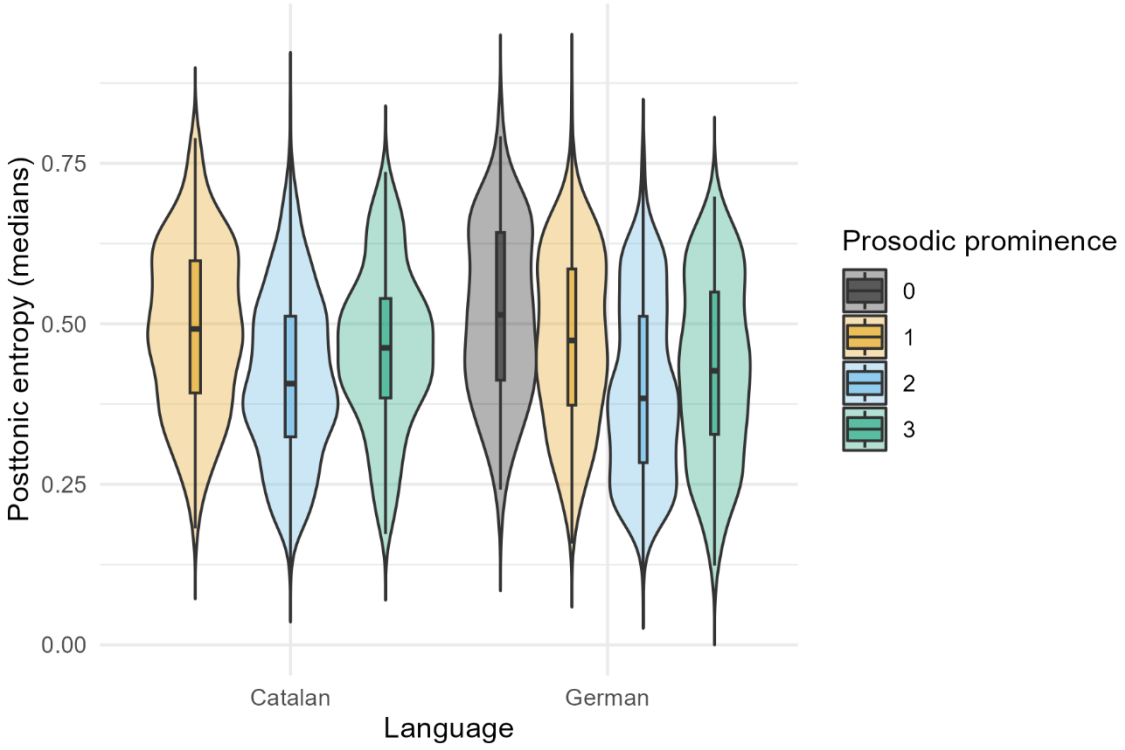
Amplitude median (without silence intervals) in the tonic syllable is higher in higher perceived prosodic prominence.

e. entropy_median (0.1735)



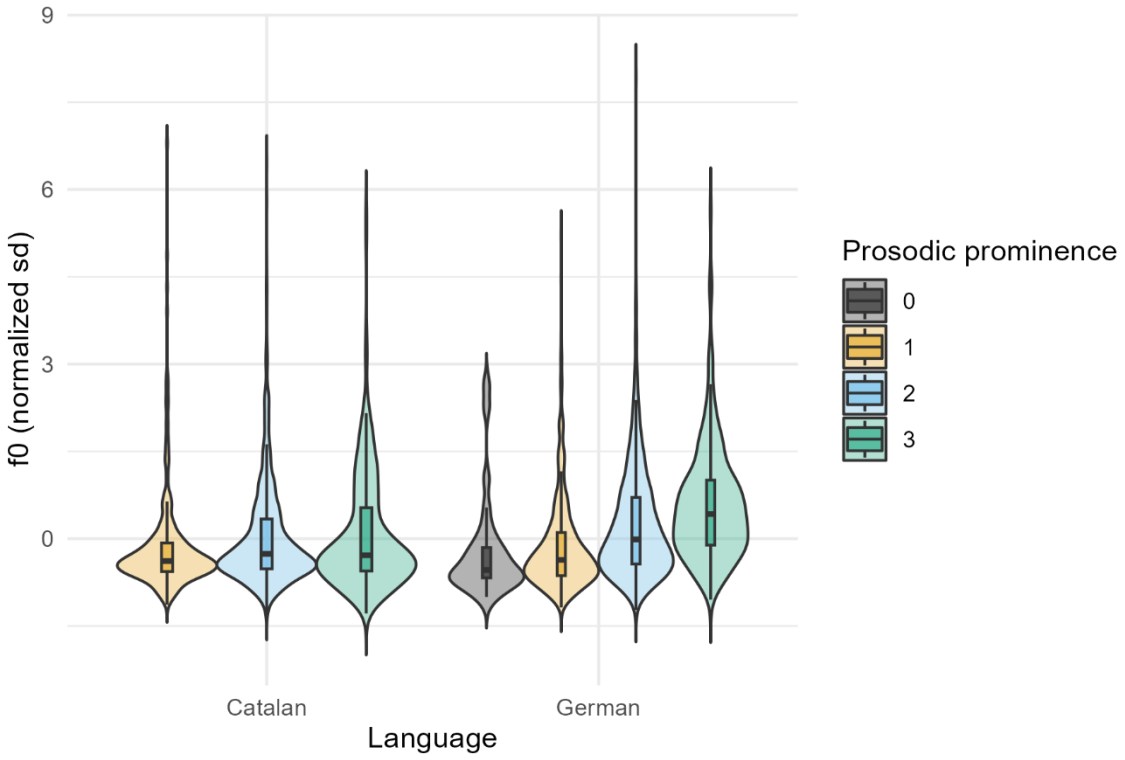
Entropy median in the tonic syllable is lower in higher perceived prosodic prominence. (Meaning that the "order" in the sound becomes higher with higher perceived prominence.)

f. entropy_medianPost (0.1698)



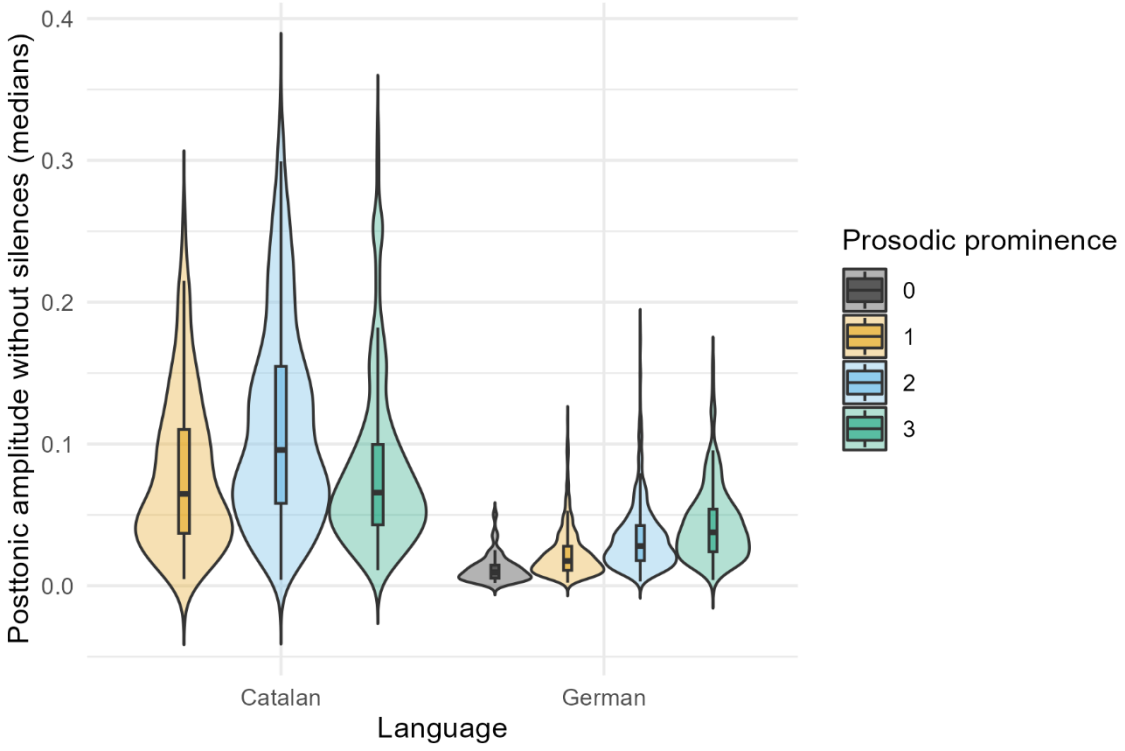
Entropy median in the posttonic syllable is lower in higher perceived prosodic prominence. (Meaning that the "order" in the sound becomes higher with higher perceived prominence.)

g. pitch_sd_norm (0.1698)



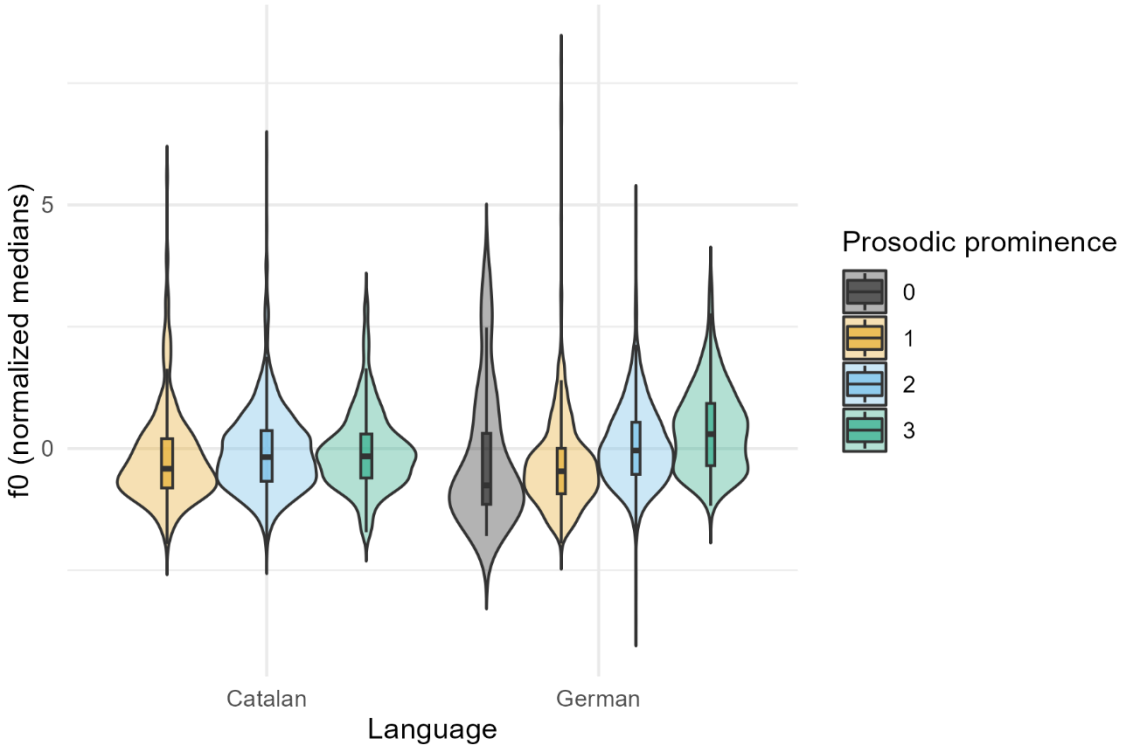
*F*0 standard deviation in the tonic syllable is higher in higher perceived prosodic prominence.

h. ampl_noSilence_medianPost (0.1612)



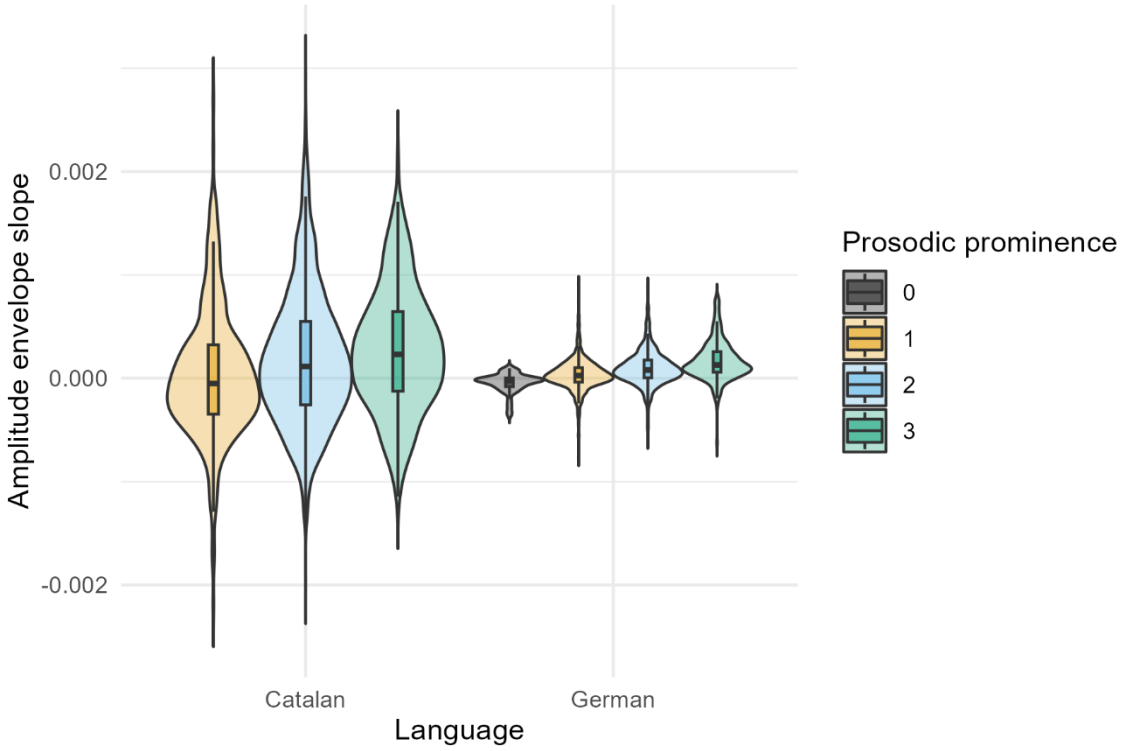
Amplitude median (without silence intervals) in the posttonic syllable is higher in higher perceived prosodic prominence.

i. pitch_median_norm (0.1581)



*F*0 median in the tonic syllable is higher in higher perceived prosodic prominence.

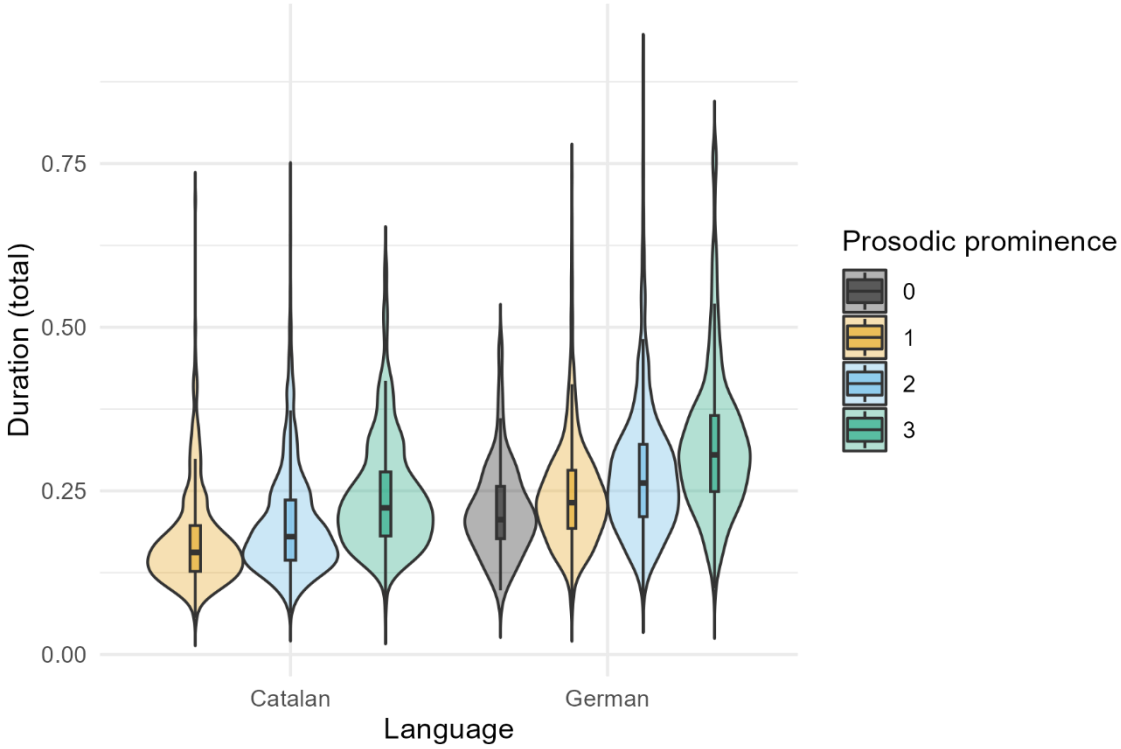
j. env_slope (0.1565)



Amplitude envelope slope in the tonic syllable is higher in higher perceived prosodic prominence.

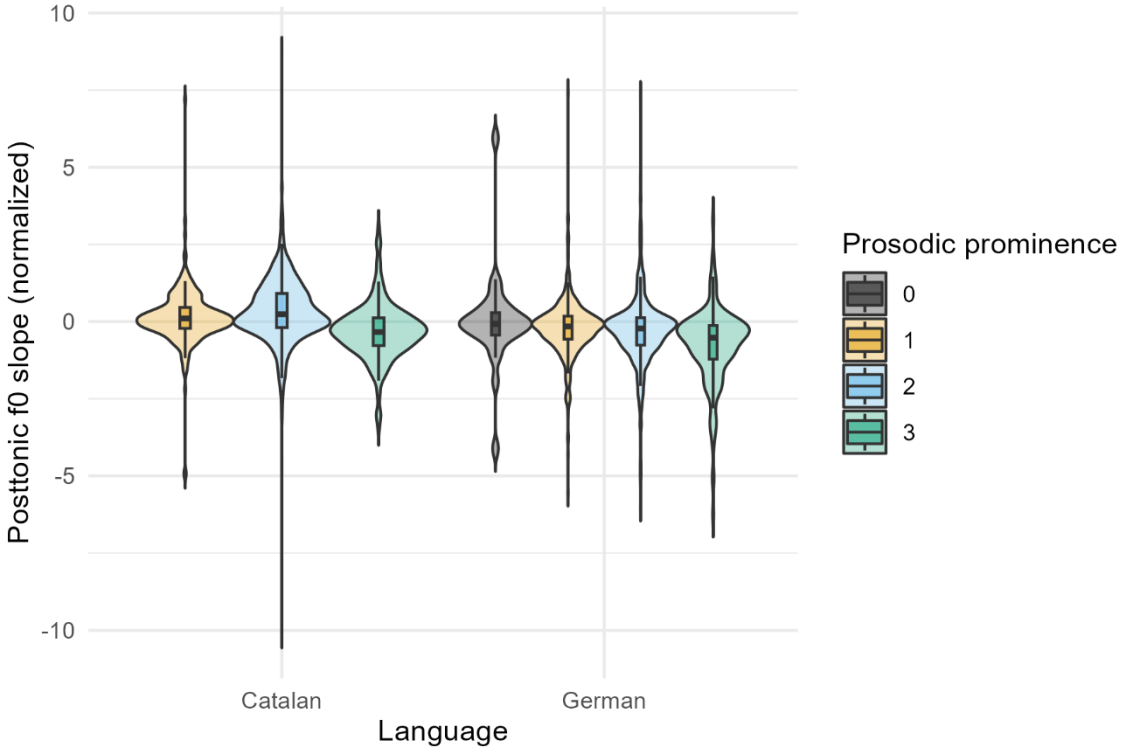
4.2. Catalan

a. duration (0.2729)



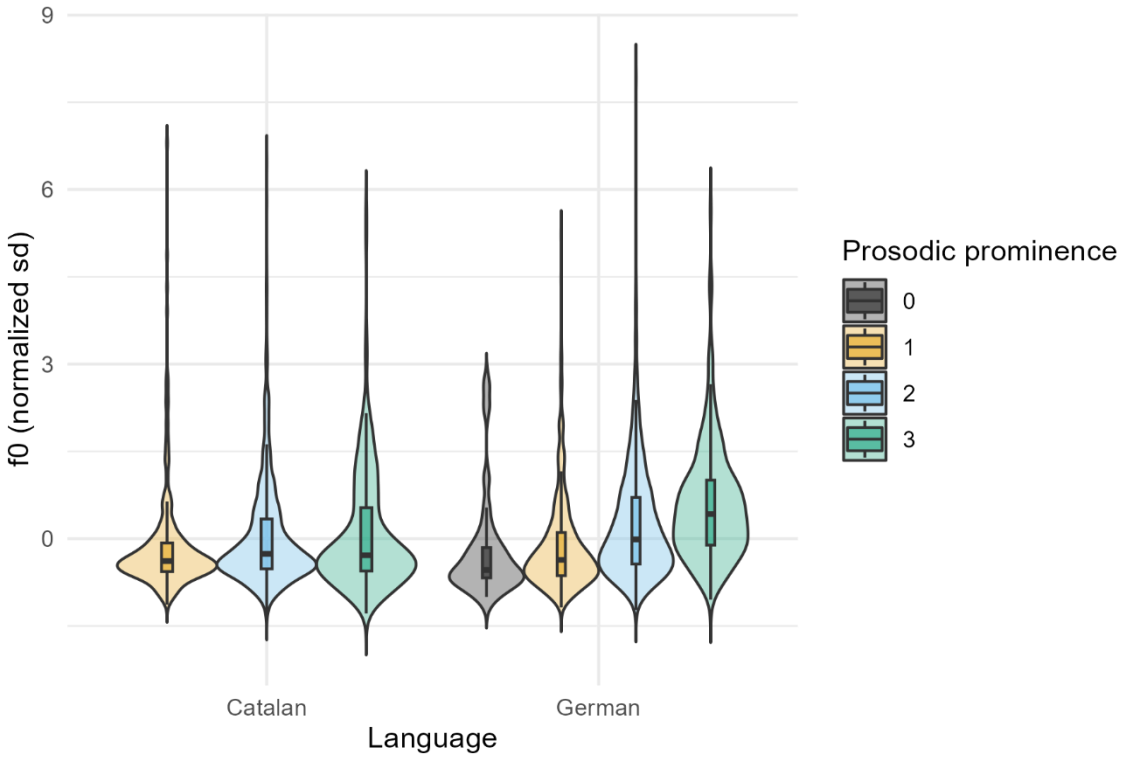
Duration in the tonic syllable is longer in higher perceived prosodic prominence.

b. f0_slope_normPost (0.1769)



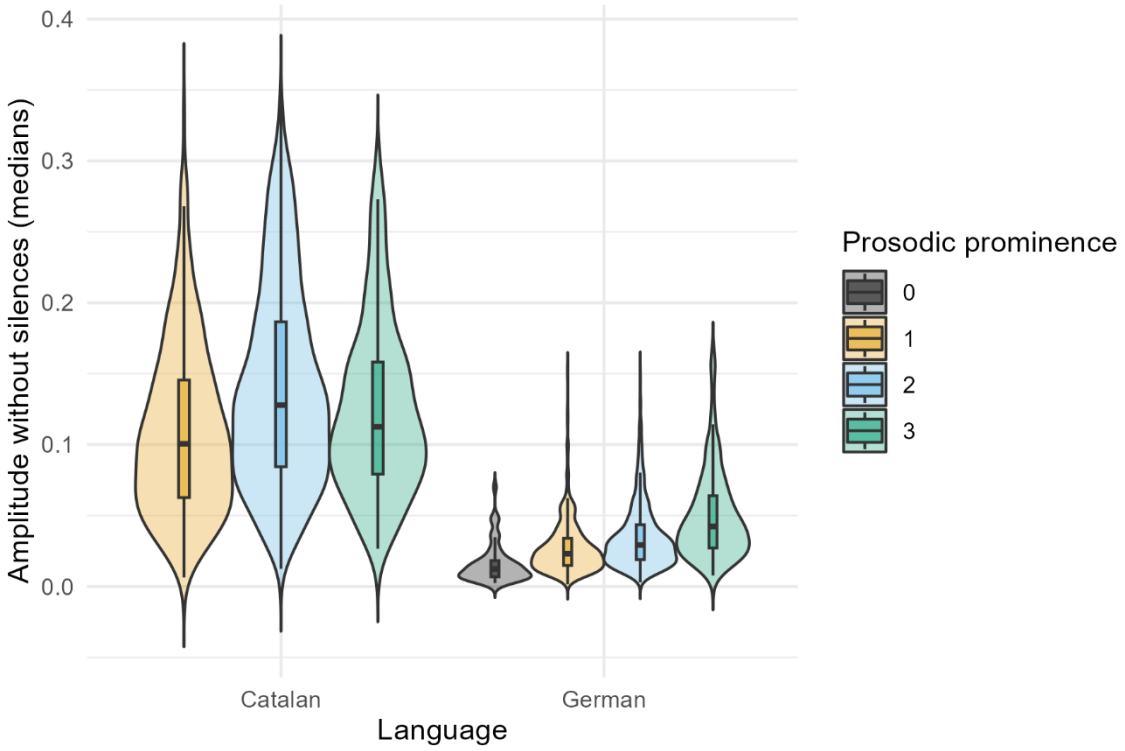
*F*0 slope in the posttonic syllable is lower in higher perceived prosodic prominence. (But the effect is not necessarily linear; it grows between 1 and 2, and falls between 2 and 3.)

c. pitch_sd_norm (0.1727)



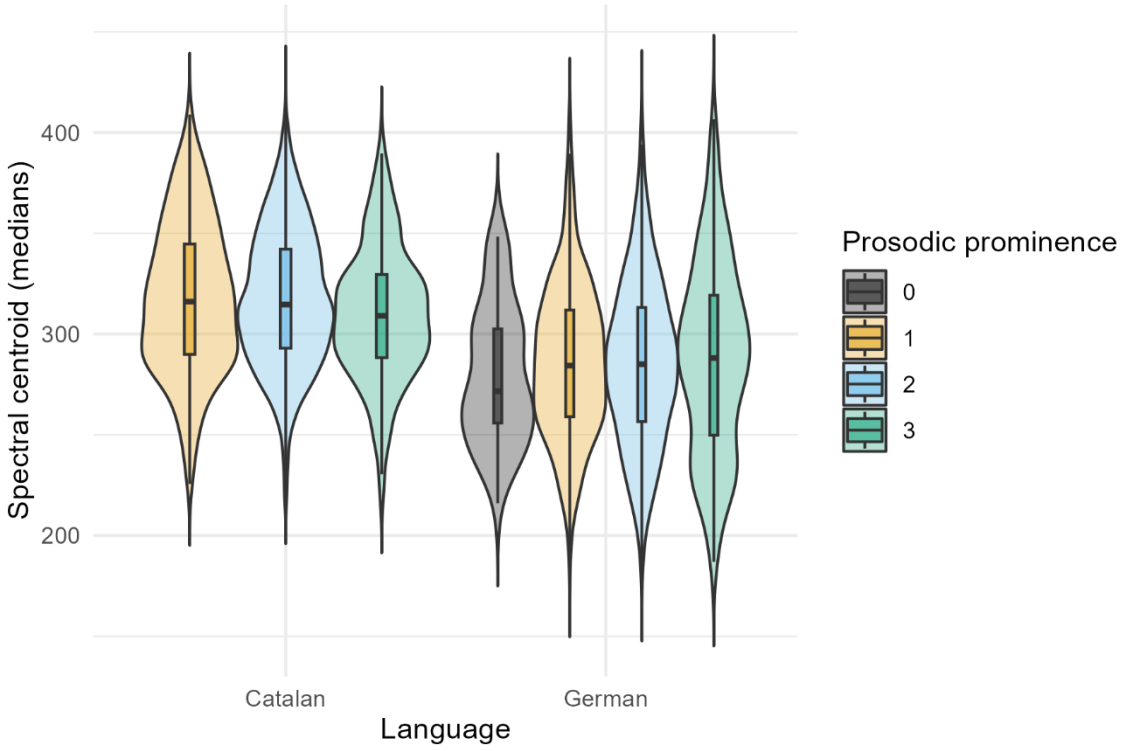
F0 standard deviation in the tonic syllable is higher in higher perceived prosodic prominence.

d. ampl_noSilence_median (0.1654)



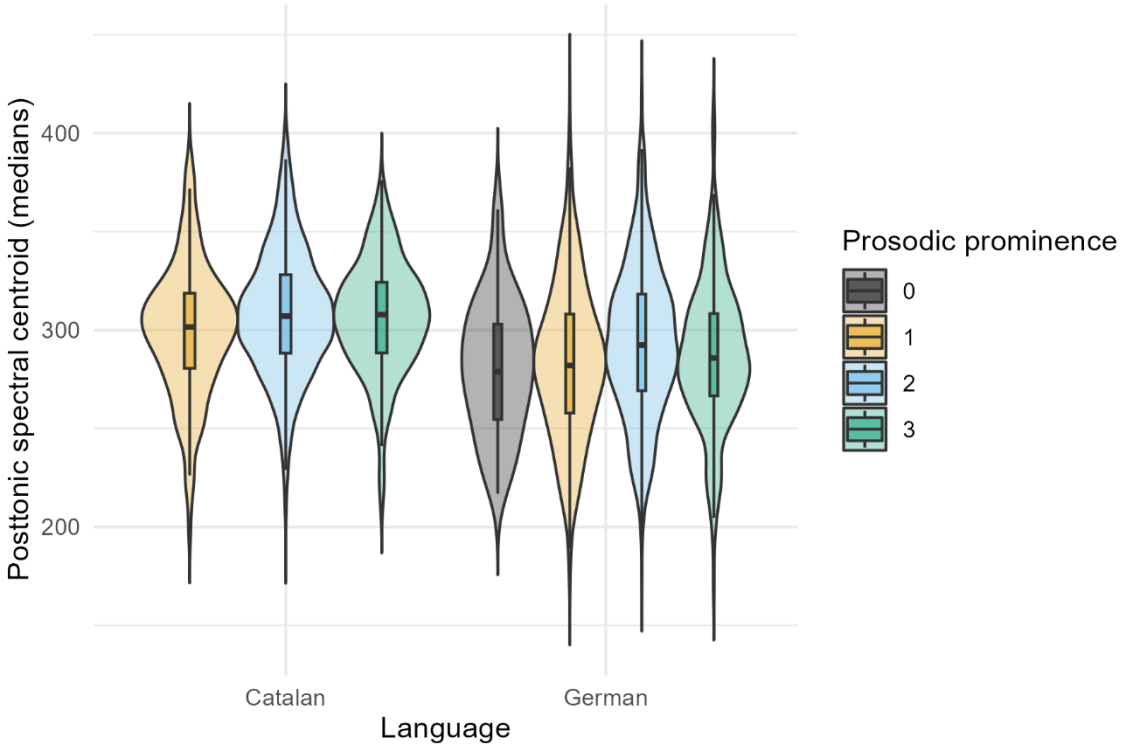
Amplitude median (without silence intervals) in the tonic syllable is higher in higher perceived prosodic prominence. (But also see that it is possibly not straightforward.)

e. specCentroid_medianPost (0.1595)



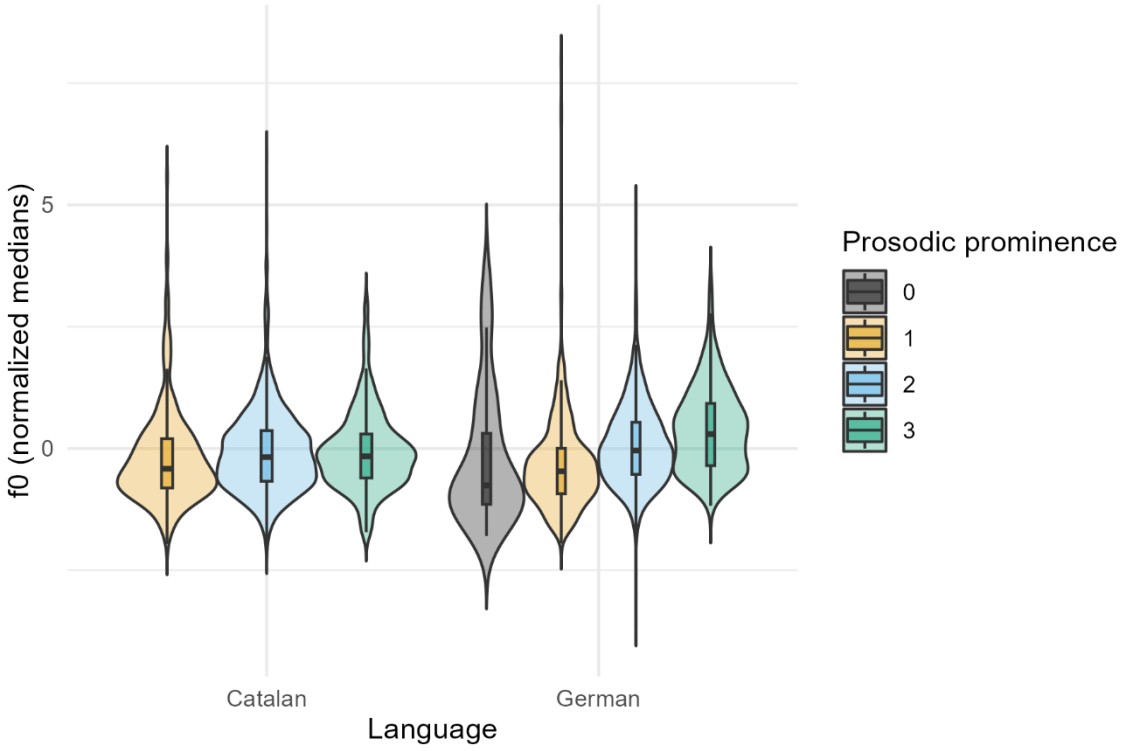
Spectral centroid in the tonic syllable is lower in higher perceived prosodic prominence. (Meaning that the sound is "darker".)

f. specCentroid_median (0.1537)



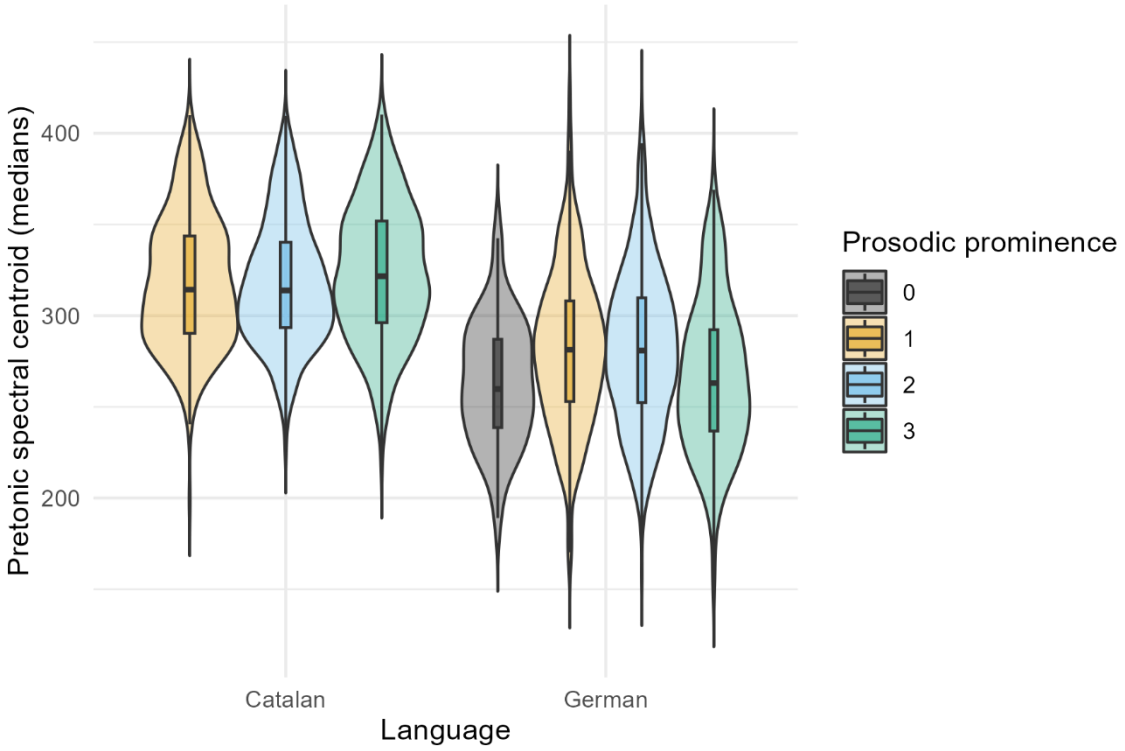
Spectral centroid in the posttonic syllable is higher in higher perceived prosodic prominence. (Meaning that the sound is "brighter".)

g. *pitch_median_norm (0.1536)*



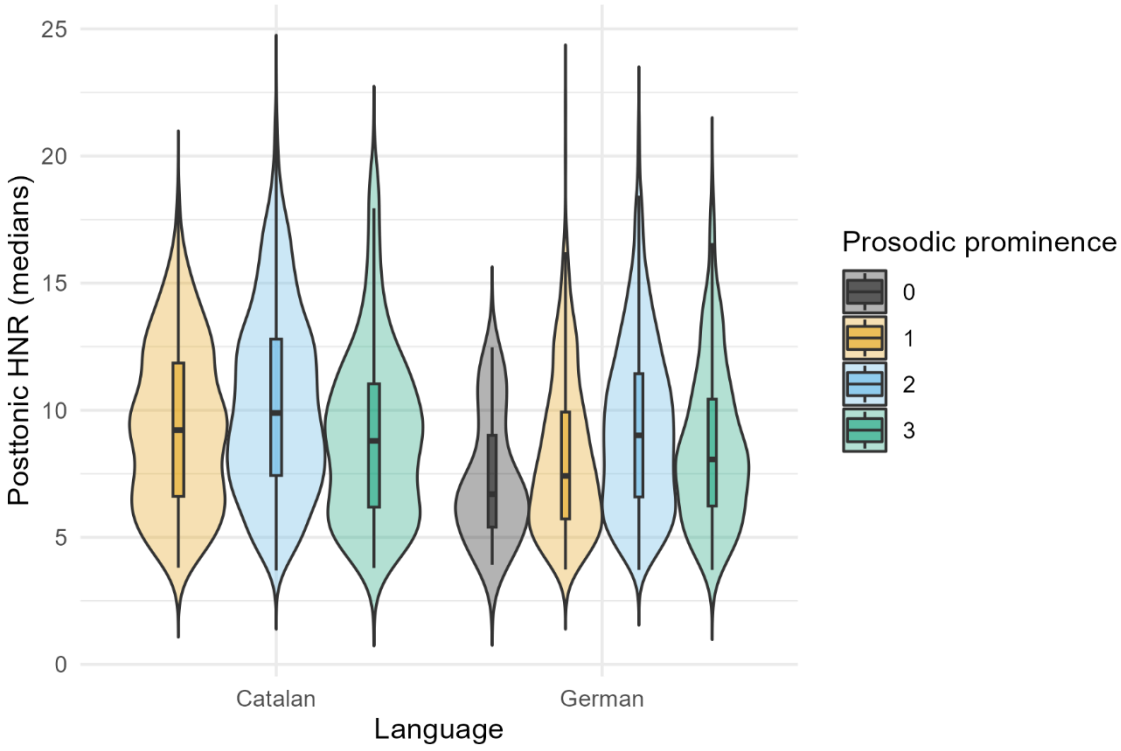
*F*0 median in the tonic syllable is higher in higher perceived prosodic prominence.

h. *specCentroid_medianPre (0.1526)*



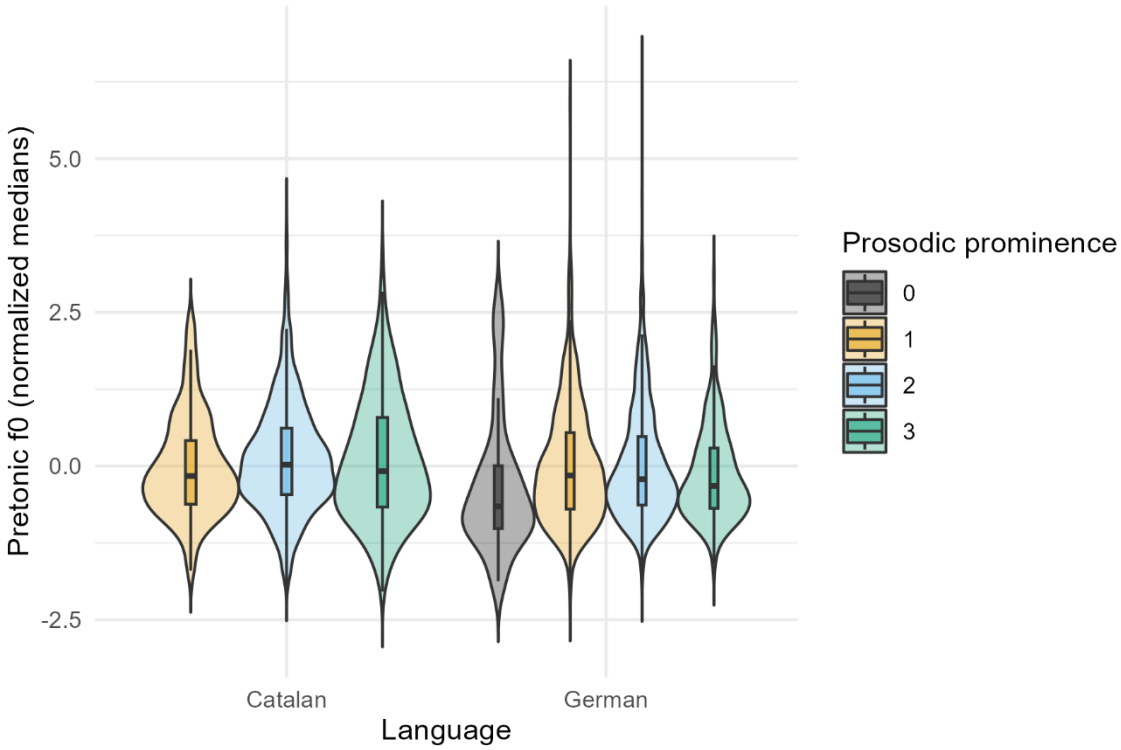
Spectral centroid in the pretonic syllable is higher in higher perceived prosodic prominence.

i. *HNR_medianPost (0.1525)*



Harmonics-to-noise ratio in the posttonic syllable is lower in higher perceived prosodic prominence. (But also not straightforward.)

j. pitch_median_normPre (0.1524)



*F*0 median in the pretonic syllable is higher in higher perceived prosodic prominence.