Supplemental Material S3. Validation of automatic methods for prosody analysis.

In this section, we present an example-based manual evaluation to validate the use of automatic extraction methods, particularly the automatic extraction of f_0; thus, in this work, we confined the "manual analysis" to the segmentation task, i.e., the use of pitch for the voiced-voiceless decision. Additionally, we included experiments with another pitch tracking algorithm (Robust Algorithm for Pitch Tracking-RAPT) to show that the errors that occurred using automatic extraction are systematic and do not significantly influence the acoustic analysis.

In the manuscript, we argued that automatic analysis can be expected to be more consistent than manual annotation, because any deviation from a supposed ground truth will be systematic and not impact the comparison. The reason is that manual analysis cannot provide a ground truth but is prone to annotator bias, especially when one has to employ several different annotators, due to the time needed for this task. Even manual segmentation has to be done based on some (ad hoc) rules—if the annotator is very experienced, this might be done in a rather systematic way. However, experience from several corpora and hundreds of hours of manual pitch segmentation tells us that we only can establish a sort of maybe "better," i.e., slightly more systematic reference. Moreover, this is only possible with only one very experienced annotator.

For instance, Batliner et al. (1993) defined six different types of laryngealizations. They are not easily told apart, and it is not easy to determine which type should be defined as voiced (i.e., with pitch) and which one as unvoiced, if at all (i.e., without pitch). This is demonstrated in Figures S2, S3, and S4, which display speech segments from the data used for the present study. Figure S2 displays a short speech segment and Figure S3 shows a section cut out from the segment displayed in Figure S2, from a female speaker in our study. Note that this speaker was chosen because she displays a relatively high portion of irregular pitch—"regular irregularities" at voiced consonants (/d/, /r/), transitions, the back vowel /a/, and "irregular irregularities" at nasals. This demonstrates that automatic pitch analysis can be pretty reliable when the pitch is regular. It is most of the time still reliable when the pitch is irregular; however, an expert would possibly segment "unvoiced" vs. "laryngealized but voiced" slightly differently. Note that this would, however, not really constitute any "ground truth" but simply another type of convention: We can define differently the periodicity needed to classify a segment as "voiced." Most important, however, is a systematic strategy. This can be expected more from the machine and less from humans—especially from different human annotators; as mentioned above, only a very experienced human expert might beat the machine. When different strategies are employed systematically to compare target and control groups, "errors" do not impact the outcome.

Figure S4 demonstrates octave errors downwards even if a human labeler would have recognized the higher pitch at ~190 Hz. The octave jumps are likely due to some "aftermath" of the previous larygealizations at /U/ and /d/ transitions. Nevertheless, we again can assume that such errors in automatic algorithms are systematic, with no pronounced bias.

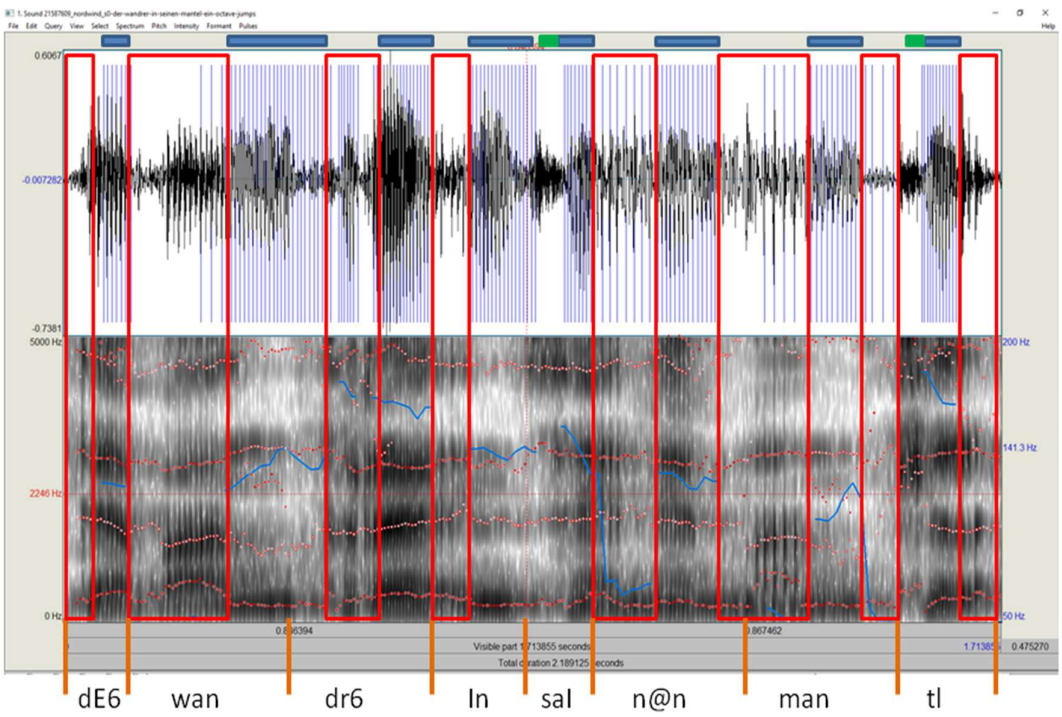

Figure S2. Excerpt from *Der Nordwind und die Sonne* (The North Wind and the Sun) *"...der Wanderer in seinen Mantel,"* SAMPA transcription. **Red boxes**: irregular phonation. **Blue bars**: reliable automatic pitch extraction. **Green bars**: voiceless segments. **Orange line**: syllable boundary. **Top signal**: time signal. **Bottom signal**: spectrogram with formants (dotted red lines) and extracted pitch curve (blue line).

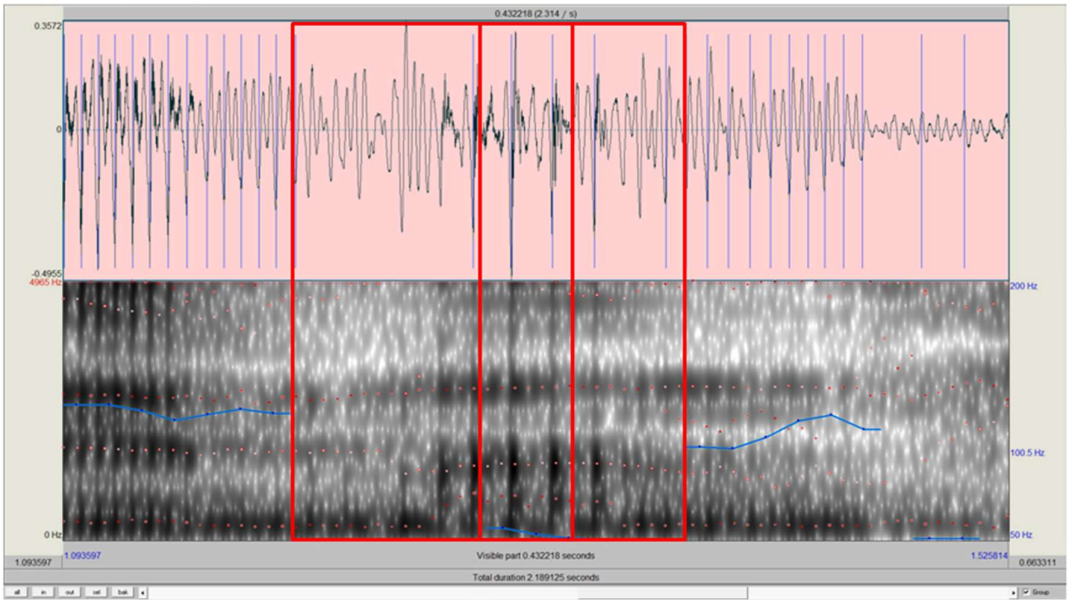

Figure S3. Zoom view of segment from Figure S2; *".. nen Man…"* **Red boxes**: more or less irregular pitch, partly automatically analyzed as voiceless, partly as very low pitch (in this case, a female speaker with ~50 Hz, i.e., laryngealized). **Top signal**: time signal. **Bottom signal**: spectrogram with extracted pitch curve (blue).

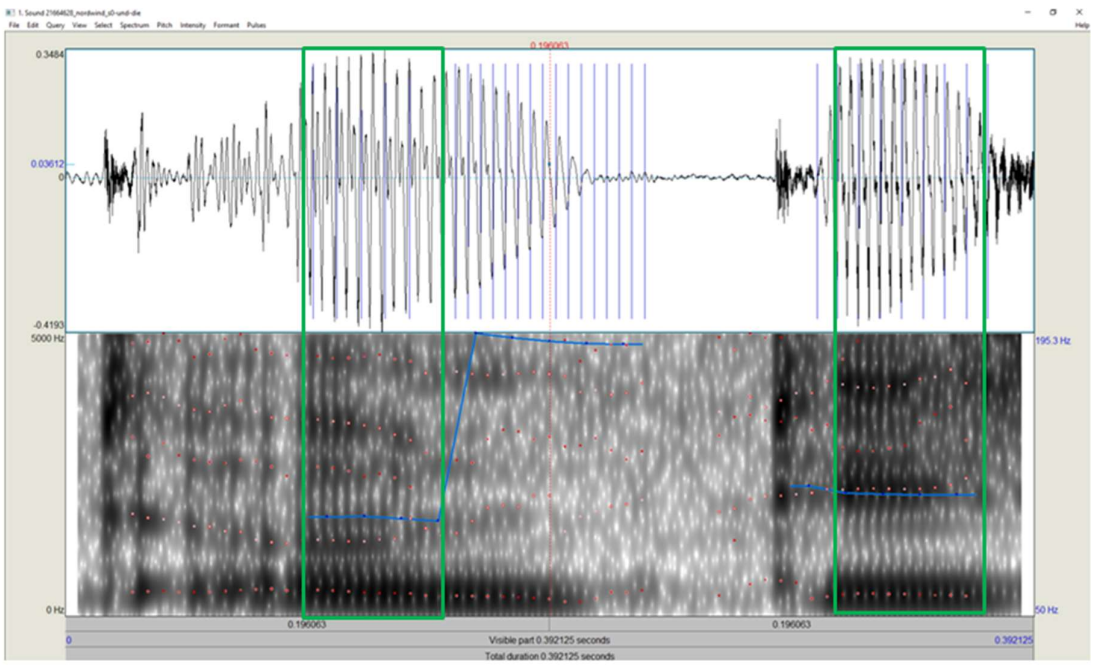

Figure S4: Octave jump errors in speech segment from the words "…und die…"; **Green boxes:** octave jump downwards. **Top signal:** time signal; **Bottom signal**: spectrogram with extracted pitch curve (blue).

f_o errors using automatic analysis.

Even though pitch doubling/halving issues occurred, the errors are systematic; thus, the results are not significantly influenced. To show that, we have performed experiments with another standard pitch algorithm called Robust Algorithm for Pitch Tracking (RAPT) to validate this argument. The results for PRAAT and RAPT can be observed in Tables S4 and S5, respectively. For both PRAAT and RAPT, the average and standard deviation of the computed f_o values are similar for CI users and controls. The *p*-values were adjusted using Benjamini-Hochberg adjustment.

Table S4. Average f_o values computed with PRAAT.

PRAAT							
Feature	Sex	CI		Controls		*p*-value	Effect size
		Mean	*SD*	Mean	*SD*		
Mean f_o [Hz]	Male	134	26	127	22	.126	0.26
Std f_o [Hz]	Male	29	8	28	8	.182	0.14
Mean f_o [Hz]	Female	193	26	195	23	.283	0.07
Std f_o [Hz]	Female	38	9	38	8	.543	0.05

Table S5. Average f_o values computed with RAPT.

RAPT							
Feature	Sex	CI		Controls		*p*-value	Effect size
		Mean	*SD*	Mean	*SD*		
Mean f_o [Hz]	Male	132	26	126	22	.142	0.26
Std f_o [Hz]	Male	26	7	25	6	.262	0.16
Mean f_o [Hz]	Female	190	26	192	24	.353	0.04
Std f_o [Hz]	Female	39	9	38	7	.309	0.13

Furthermore, we compared the percentage of f_o outliers outside the interquartile range (IQR) for the two different pitch tracking algorithms. Figures S5 and S6 show the results. Upward and downward outliers refer to f_o values outside the IQR (and whiskers). Furthermore, the percentage of f_o errors was computed for PRAAT and RAPT. The percentage of errors is relatively low for both pitch tracking algorithms. Furthermore, the difference between CI users and controls is consistent for PRAAT and RAPT.

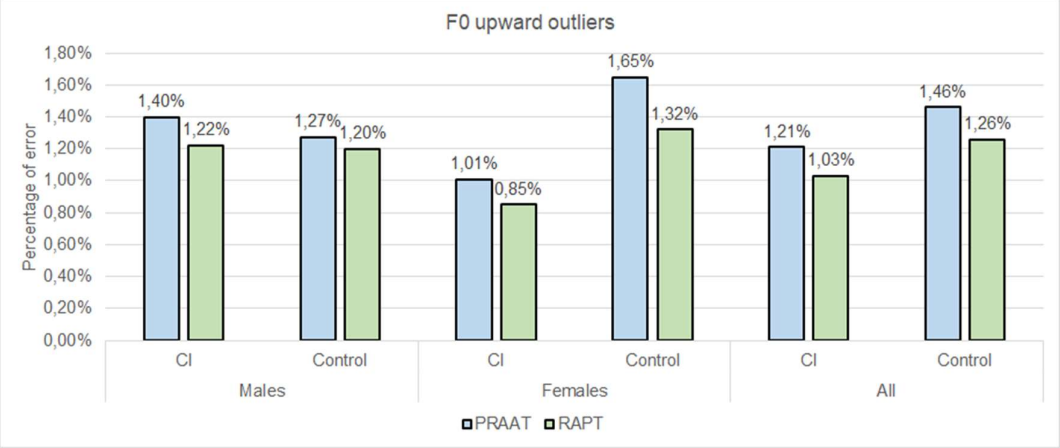

Figure S5. f_o errors outside the upper whisker (third quartile + 1.5 × IQR). **Blue bars:** f_o values obtained with PRAAT. **Green bars:** f_o values obtained with RAPT.

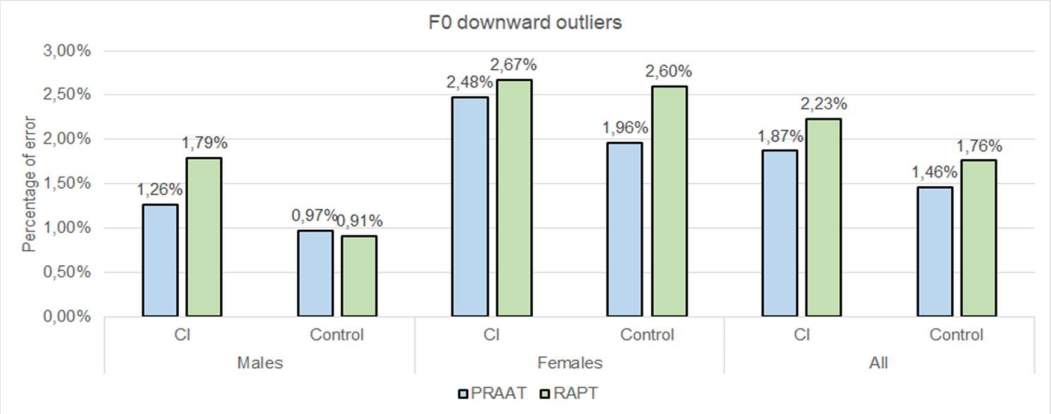

Figure S6. f_o errors outside the lower whisker (first quartile – 1.5 × IQR). **Blue bars:** f_o values obtained with PRAAT. **Green bars:** f_o values obtained with RAPT.

To summarize: Own experience based on extensive research tells us that there will be no falsifying differences between manual and automatic processing—given the quality of recordings and the sample. This is, of course, no proof. However, verifying this with a small sample would not be watertight; verifying this with a large sample would require many hours of manual work and had to be based on criteria that are not 'objective' but partly arbitrary.

Note that we deal with the identical read text; thus, differences in voiced-unvoiced cannot be caused by different texts. However, when a specific sub-group displays more unvoiced parts, this can be traced back to more irregular phonation—automatically analyzed either as octave-jump or unvoiced, i.e., no pitch value.

The effect size we are looking for interpreting is large enough so we can be positive that we do not run into spurious effects. Moreover, it can be assumed that algorithmic biases are more systematic than human biases, especially among different human annotators.