Supplemental Material S1. Visual sort-and-rate (VSR) task for synthetic quality.

Additional VSR training and experimental tasks were performed at the end of the session to determine how synthetic the synthesized stimuli were perceived. This task was included to ensure that potential differences in strain ratings between RFF-modified and unmodified samples or between samples with and without noise were not due to the modified samples sounding synthetic. Participants completed a VSR training module before the actual VSR task for rating synthetic quality. We presented the same eight samples that contained a wide range of strain and were used in the training module before the VSR task for rating strain. We placed all of them at 0 on a synthetic quality scale, and listeners were informed that these samples were rated 0 for synthetic quality because all of them were produced naturally (i.e., objectively not synthetic) and were not synthesized, despite having various levels of strain.

The same protocol from the VSR task for strain ratings was provided to the listeners to complete the experimental VSR module for rating synthetic quality. Each set was designed specifically for each listener as described in the methods and a total of 10 sets containing 80 items (64 stimuli + 16 stimuli for intrarater reliability) were completed. The VSR task for synthetic ratings took approximately 15 minutes to complete. Synthetic quality ratings for each stimulus obtained from the VSR tasks were averaged across the listeners. A three-way repeated measures ANOVA was performed on mean synthetic quality ratings. Synthetic quality showed intra-rater reliability (Pearson's r) above .7 in only 11 of 20 listeners (median = .71, range = –.12 – .91) and poor interrater reliability (ICC = 0.18, 95% CI = 0.13 – 0.26).

The effect of RFF modification in the three-way ANOVA on synthetic quality was not statistically significant (p = .67), which indicates that the synthetic quality ratings between the samples with and without RFF modification were not statistically different (Figure S1). Thus, the statistically significant effect of the interaction between vocal effort level and RFF modification on strain is unlikely to have resulted from the RFF-modified samples possibly sounding synthetic. However, the effect of noise on synthetic quality was statistically significant and had a large effect size (p = .001, η_p^2 = .81). The samples with added noise had increased synthetic quality ratings (Figure S1, below), which could suggest that increased synthetic quality in the samples with noise might have contributed to their increased strain ratings.

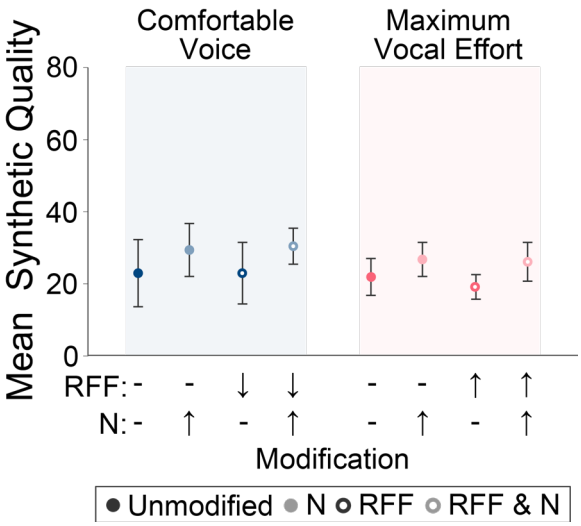

Figure S1. Mean synthetic quality ratings of comfortable and maximum effort samples as a function of modification condition. There were no statistically significant differences among conditions ($p > .05$). Error bars indicate 95% confidence intervals. (abbreviations: RFF = relative fundamental frequency, N = noise, - = unmodified, ↑ = increase, ↓ = decrease).