Comparing the imitation of naturally-produced and synthesized F0 in American English nuclear tunes

Jeremy Steffman[1], *Jennifer Cole*[2]

[1]The University of Edinburgh
[2]Northwestern University

jeremy.steffman@ed.ac.uk, jennifer.cole1@northwestern.edu

Abstract

Imitation tasks are used in intonation research to identify properties that are perceptually salient and encoded for subsequent production. The current study examines whether and how imitation of synthetic versus naturally produced F0 contours may differ. We compared F0 contours in American English from two imitation experiments where participants heard sentences with the same phrase-final intonation and reproduced the heard pattern on a novel sentence. In one experiment, F0 patterns of stimuli were controlled via pitch resynthesis using straight-line approximations of (phonological) tonal targets; the other used natural productions of the same tunes. F0 trajectories were examined to identify which F0 properties of the stimuli were preserved or lost as a function of the type of stimulus. Imitations of natural vs. resynthesized stimuli were compared using time-series k-means clustering analysis, GAMM modeling, and RMSD as a measure of F0 difference between imitation and stimulus. We observe striking similarity in imitations of natural and resynthesized stimuli based on clustering solutions, with small, localized differences in GAMMs for only two out of eight tunes tested. RMSD results show closer imitation with resynthesized stimuli, suggesting greater attention to fine phonetic detail of F0 patterning when other cues to intonational contrasts are held constant.

Index Terms: intonation, imitation, nuclear tunes

1. Introduction

One core challenge that intonation researchers are faced with is how to elicit particular intonational tunes from naive speakers. A solution that has a long history in the field is imitation [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]: participants are exposed to a stimulus and asked to reproduce what they have heard. The imitated productions can be used to infer what aspects of the stimuli are important to listeners, and to inform theories of intonational phonology and speakers' mental representation of intonational tunes.

The premise of the current study was to assess a basic empirical question in a relatively controlled fashion: how do speakers imitations of resynthesized F0 trajectories compare to imitations of naturally produced tunes? (Re)synthesizing F0 ensures control over F0 contours. However, this also comes with a clear downside: resynthesized F0 may only approximate naturally produced contours, and depending on the nature of the resynthesis may lack features that are present in natural intonational melodies. For example, straight-line F0 approximations are a common way to represent tonal targets and their sequencing [11], but these do not capture variation in F0 rise/fall shape which are produced by speakers [12]. Conversely, with naturally-produced stimuli, a great deal of control may be sacrificed, which, depending on the research question, may make use of natural stimuli untenable. If F0 is the cue of interest, resynthesis offers a way to vary it in isolation.

In the two studies we report here, the original goal was to assess the distinctiveness of nuclear tunes in American English, as imitated by native speakers. We resynthesized eight nuclear tunes, tonally specified with a pitch accent {H*,L*}, phrase accent {H-,L-}, and boundary tone {H%,L %}. We compared two studies, both of which are variants of the same paradigm, testing the same nuclear tunes. In one study, the stimuli for imitation were created by re-synthesizing F0. In the other, naturally produced imitations (of the same tunes) from a previous experiment were used as stimuli. This is an exploratory study in which we performed bottom-up clustering analysis [13] on the imitated F0 trajectories to assess which distinctions emerge from each experiment. In addition, we use a GAMM to model variation in F0 trajectories conditioned by the tune label of the stimulus that was the intended target of imitation, with data pooled across the two experiments. GAMM output is assessed for differences based on the stimulus type, examining pairwise comparisons of difference smooths. Finally, we asses the "closeness" of imitated tunes to model stimuli, to see if speakers more accurately reproduce the F0 pattern of the stimulus for one type of stimulus compared to the other.

2. Methods

We recruited 30 self-reported monolingual speakers of American English for each of the two experiments (60 speakers total, with no overlap between experiments). In both we analyze just the nuclear tune, produced over trisyllabic stress-initial words.[1]

2.1. Imitative speech production experiments

The first experiment, referred to as the resynthesized stimuli experiment, is essentially a replication of [2], in which resynthesized stimuli were played to participants. Two model speakers (one male, one female) were recorded and the stimuli consisted of two sentences: "Her name is Marilyn" and "He answered Jeremy". The nuclear word was resynthesized to have one of eight tunes, using [14, 15]. We represent these eight tunes without the diacritics, using HHH instead of H*H-H%, and so on. Stimuli were based on straight-line approximations from [16, 11], as shown in Figure 1, panel A.[2] On a given trial, a participant heard two model stimuli, each with the same tune, produced by the two speakers, with two different sentences (this

[1]Stimuli are available in an open access OSF repository at: https://osf.io/32fwv/.

[2]The reader is referred to [2] for more details on stimulus creation (the present study used two of the three model sentences in that study).

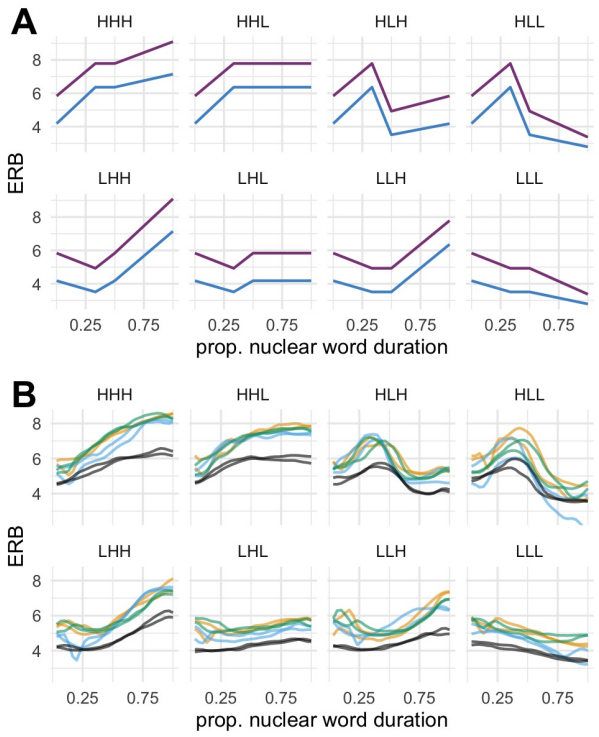

Figure 1: *Resynthesized (panel A) and naturally produced (panel B) stimuli F0, shown over normalized time, in ERB. Tunes are shown in the panels of the plot and the speaker identity is indicated by line color (two speakers in panel A, four in panel B).*

2×2 speaker×sentence combination resulting in four unique combinations per tune). The participant was prompted to produce the tune model over one of two new sentences "She gathered lavender" and "They honored Melanie". This target sentence, crossed with the eight tunes and four stimulus conditions yielded 128 trials. These were supplemented with 16 additional trials that did not fully cross the conditions but elicited each tune twice more with variation in stimulus condition and target sentence, for a total of 144 trials.

The second experiment, which will be referred to as the natural stimuli experiment, used as stimuli select imitated productions from four speakers (two male, two female) who were participants in [2] for stimuli. The natural stimuli are shown in Figure 1 (panel B). These speakers were selected as those who produced clear imitations of each of the eight tunes, on the basis of an inspection of their F0 trajectories. Our aim was to select speakers who seemed to best capture the distinctions present in the stimuli in order to compare natural and resynthesized productions in a more direct way. In a given trial in this experiment, we opted to present just one speaker's voice, producing one of two sentences: "She remained with Madelyn" and "He modeled harmony". The target sentences were "She gathered lavender" and "They honored Melanie". There were a total of 128 trials which crossed model speaker with target sentence, the order of model sentences, and tune (4×2×2×8).

There are several differences between the experiments which should be kept in mind. In the resynthesized experiment, both model speakers were heard on a given trial, while only one model speaker was heard in a trial in the natural experiment. In addition, stimuli from four speakers were used in the natural experiment, while source recordings from only two speakers were used for the resynthesized experiment. Both experiments do share the critical tunes and the fact that two model stimuli were heard on a given trial. In both experiments we measured F0 using STRAIGHT, in VoiceSauce [17, 18]. F0-tracking errors were assessed using [19], which flags files as containing potential errors when they contain sudden jumps in F0. Flagged files were inspected to confirm the presence of an error. In total, this led to the exclusion of 10% of the files in the resynthesized experiment (3,848 retained for analysis) and 13% in the (3,332 retained for analysis). Time-normalized F0 trajectories were output with 30 samples per nuclear word. We carried out clustering analyses [13] (described in detail below) on both sets of stimuli (not shown), which confirmed that the eight nuclear tunes were well separated in both cases; for both sets of stimuli, eight was the optimal number of clusters, which cleanly separated the eight stimulus tunes.

2.2. Analyses

Clustering: To assess emergent distinctions in the imitations of naturally-produced and resynthesized tunes we used time-series k-means clustering [13]. The input time-normalized trajectories were computed as speaker means for each tune, and scaled within speaker to normalize for differences in speaker F0. As such, each speaker contributed eight trajectories, with 240 input for each experiment (30×8). Here we report the optimal partition of the input trajectories, which is the number of clusters that results in an optimal ratio of within-cluster to between-cluster dispersion, calculated using the Caliński-Harabasz criterion [20]. Our interest in assessing the two clustering solutions will be in determining if and how they vary across experiments, and the ways in which exposure tunes (shown in Figure 1) map to the emergent clusters.

GAMM modeling: We fit a GAMM to the data [21, 22], modeling F0 over time, as a function of tune, and of stimulus type. The GAMM was fit with a combined tune + stimulus type variable. This was included as both a parametric term, and smooth term over time in the model. Our focus in inspecting the model fit was then to compare resynthesized to natural stimuli, within a given tune. The models were fit using the default basis functions and k specification for smooth terms which was determined to be adequate using the gam.check function. Random smooth were fit using the reference-difference method [23] for speaker by tune.

RMSD: To assess the extent to which speakers' productions were close to the F0 in model tunes, we computed root mean squared distance (RMSD) between a speakers' production on a given trial, and the immediately preceding model stimulus (i.e., the second stimulus heard on that trial). Both the model stimulus and the speaker's production were represented as time-normalized contours with 30 samples over the nuclear word. We opted to compute both as speaker-centered, but not scaled, F0 measured in ERB. [2] has shown previously the imitations tend not to match model speakers' pitch height, but instead reflect relative changes in F0 within a speaker's pitch range. Thus by centering F0 we are capturing how much imitated F0 follows the model speakers in terms of deviations from speakers' mean F0. This still allows for the capturing of differences in F0 range across models and participants (though not F0 height), where for example, if a participant does not produce as large of an F0 excursion as a model speaker in a given trial, they will have higher RMSD. We modeled this variable using a Bayesian lin-

ear regression model [24], and used [25] to compute marginal contrasts. The model was fit to predict RMSD as a function of model tune, experiment (resynthesized versus natural), and the interaction of these two fixed effects. Random effects were a random intercept for speaker (participant), with a by-speaker slope for tune, as well as a random intercept for model speaker (one of six, two being resynthesized, four being natural). The model was fit using a log-normal family with fixed effect priors and intercept set to be normal(0,1) in logged RMSD values. In reporting the results from the model we give estimates for a fixed effect or marginal mean ($\hat{\beta}$) and 95% credible intervals (CrI). When these intervals exclude the value of zero we take this to be robust evidence for an effect or computed marginal contrast. To quantify the strength of evidence for a given effect we also report the percentage of a given posterior estimate which has a particular sign, computed with [26]; when this metric exceeds 97.5% this corresponds to credible intervals excluding 0, though values approaching this can be taken as graded evidence for an effect, this will be indicated as "pd" (probability of direction).

3. Results and conclusion

Clustering analysis: Figure 2 shows the results of the two clustering analyses. For both, evaluation of 2 through 8 clusters determined 5 to be the optimal number of clusters. The clustering partition of the data is highly similar in both cases to what was observed in [2]. In the case of the resynthesized stimuli, this is perhaps not surprising, given that the stimuli are the same (the only difference being that listeners heard two model stimuli in the imitations of resynthesized tunes presented here, and heard three in [2]). The breakdown is, roughly, that {HHH,HHL} are partitioned into a single high-rising cluster A {HLH,HLL} are partitioned into a single rising-falling cluster B, and {LHL,LLH} into a low-to-mid rising cluster C. LLL and LHH are overall well-separated and partitioned into clusters D and E. Given that both clustering partitions selected the same number of clusters as optimal, with roughly the same tune-to-cluster mapping, we do not have evidence for substantial differences in the imitations of resynthesized versus natural tunes. We thus conclude that the two groups produce the same number of robustly distinct F0 shapes, which are related to the input tunes in a similar manner. One qualitative difference can be noted: LLH is somewhat split across clusters C and D in imitations of the natural tunes, and is more directly mapped to cluster C in imitations of resynthesized tunes. This could be due to increased similarity between LLH and LHH in the natural stimuli.

GAMM analysis: Reflecting the relative similarity between clustering solutions, the GAMM model found minimal differences for the imitations of natural or resynthesized versions of each of the eight tunes.As shown in Figure 3 Panel A, there were two (of eight) tunes for which a significant difference between experiments was detected in some region of the F0 contour. The first is LHH, for which the F0 scaling at the end of the contour is overall lower in the natural experiment. The second is LLH, for which there is a window in the middle-end of the contour for which the resynthesized experiment has overall lower F0. Both are shown in light-gray highlighting. For all other tunes, there was no significant difference across experiments according the GAMM. The observed difference for LHH comports with the clustering results in the sense that imitations of LLH in the natural data get divided between cluster C (the LLH and LHL cluster) and cluster D (the LHH cluster), which is likely due to the lower trajectory for LHH in the natural data. Otherwise, the GAMM results lead us to conclude that the imitations of natural or resynthesized stimuli are remarkably similar in terms of the shape of the imitated contours. One important consideration is that the GAMM assessment is fundamentally shape-based in the sense that we analyzed scaled F0 values, which, by scaling within speaker, represent F0 values are relative excursions from the speaker's mean, such that any scaling differences across speakers (i.e. different pitch ranges)

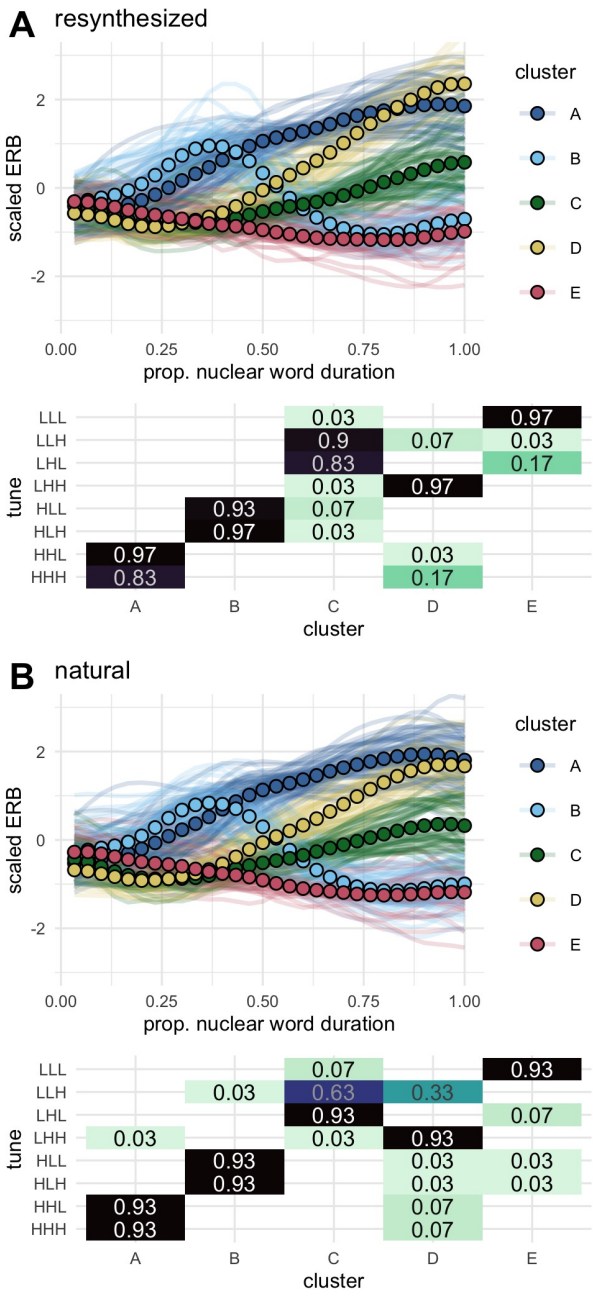

Figure 2: *k-means clustering solutions for the imitations of resynthesized tunes (A) and naturally produced tunes (B). Each panel shows the trajectory means for each cluster (dots) and contributing trajectories (lines). The heat maps below show the proportion of each of the eight tunes (rows) which were placed in each cluster (columns).*

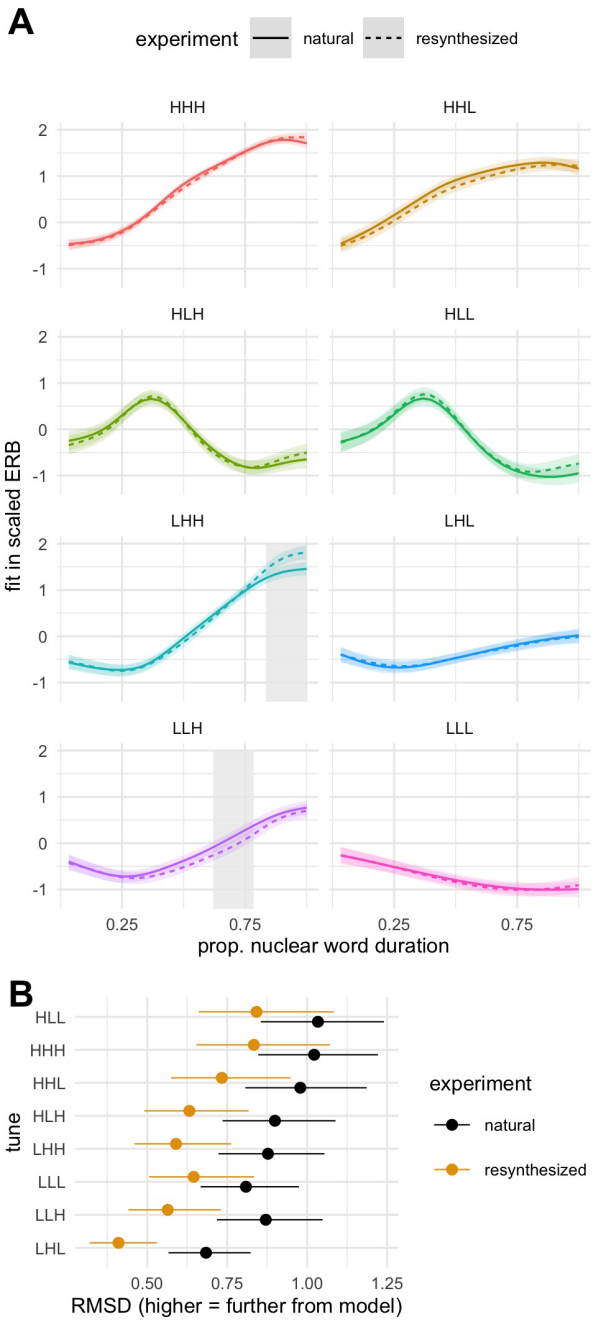

Figure 3: *Panel A: GAMM fits to the data showing 95% CI. Regions of significant difference for a given tune, as assessed by the inspection of pairwise difference smooths are shown by light gray shading. Panel B: Model estimates for RMSD based on experiment and tune. Points show posterior medians and error bars show 95% CrI.*

are not captured.

RMSD analysis: Model RMSD estimates for each tune and experiment are shown in Figure 3 Panel B. The marginal effect comparing natural and resynthesized experiments overall found that RMSD was higher (i.e., participants were less like the model stimuli) in the natural experiment as compared to the resynthesized experiment ($\hat{\beta}$= 0.29, 95%CrI = [0.02,0.56]). There were also differences among tunes, which were overall fairly comparable in terms of their rank order in both experiments. This is shown in Figure 3B, where RMSD is plotted by tune and experiment, which is sorted based on the RMSD for a given tune, aggregated by experiment. For three tunes the credible intervals for the marginal effect of experiment included zero, though the posterior showed a clear skew, hence weaker evidence for an effect. These were: HHH (pd = 92), LLL(pd = 92), and HHL(pd = 96). For the other five tunes, there was clear evidence for increased closeness to the model in the resynthesized experiment.

In conclusion, in this study we set out to examine a methodological question pertaining to the imitation of intonational tunes. We found strikingly similar emergent clusters (in both cluster number and composition), and reproduced F0 trajectory shapes in GAMM modeling. However, we find that when only F0 is the only acoustic tune correlate that varies, and F0 variation is highly controlled, speakers' imitations are closer to the model. From this perspective, though the similarity in reproduced shapes is rather striking, we suggest that F0-only (resynthesized) imitation may be more suited to examining the extent to which particular patterns are veridically/accurately reproduced. If however, the imitations themselves (alone) are the object of inquiry then the comparisons outlined here suggest that speakers reproduce similar F0 distinctions in the imitation of (re)synthesized or naturally-produced tunes. Future work may benefit from considering other techniques for F0 synthesis, including those that are not straight-line approximations.

4. Acknowledgments

We are thankful to Chun Chan for technical help. This work was supported by NSF BCS-1944773 (Cole, PI).

5. References

[1] B. Braun, G. Kochanski, E. Grabe, and B. S. Rosner, "Evidence for attractors in English intonation," *The Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 4006–4015, 2006.

[2] J. Cole, J. Steffman, S. Shattuck-Hufnagel, and S. Tilson, "Hierarchical distinctions in the production and perception of nuclear tunes in American English," *Laboratory Phonology*, vol. 14, no. 1, 2023.

[3] L. C. Dilley, "Pitch range variation in english tonal contrasts: Continuous or categorical?" *Phonetica*, vol. 67, no. 1-2, pp. 63–81, 2010.

[4] L. C. Dilley and C. C. Heffner, "The role of F0 alignment in distinguishing intonation categories: evidence from American English," *Journal of Speech Sciences*, vol. 3, no. 1, pp. 3–67, 2013.

[5] C. Gussenhoven, "Experimental approaches to establishing discreteness of intonational contrasts," *Methods in empirical prosody research*, pp. 321–334, 2006.

[6] C. Petrone, D. d'Alessandro, and S. Falk, "Working memory differences in prosodic imitation," *Journal of Phonetics*, vol. 89, p. 101100, 2021.

[7] J. B. Pierrehumbert and S. A. Steele, "Categories of tonal alignment in English," *Phonetica*, vol. 46, no. 4, pp. 181–196, 1989.

[8] J. Steffman, S. Shattuck-Hufnagel, and J. Cole, "The rise and fall of American English pitch accents: Evidence from an imitation study of rising nuclear tunes," *Proc. Speech Prosody 2022*, pp. 857–861, 2022.

[9] J. Steffman, J. Cole, and S. Shattuck-Hufnagel, "Intonational categories and continua in American English rising nuclear tunes," *Journal of Phonetics*, vol. 104, p. 101310, 2023.

[10] K. Zahner-Ritter, M. Einfeldt, D. Wochner, A. James, N. Dehé, and B. Braun, "Three kinds of rising-falling contours in German wh-questions: Evidence from form and function," *Frontiers in Communication*, p. 58, 2022.

[11] N. Veilleux, S. Shattuck-Hufnagel, and A. Brugos, "Transcribing prosodic structure of spoken utterances with tobi (version 6.911)," 2006, Massachusetts Institute of Technology: MIT OpenCourseWare. [Online]. Available: https://ocw.mit.edu

[12] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, "Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology," *Laboratory Phonology*, vol. 3, no. 2, pp. 337–383, 2012.

[13] C. Genolini and B. Falissard, "Kml: A package to cluster longitudinal data," *Computer methods and programs in biomedicine*, vol. 104, no. 3, pp. e112–e121, 2011.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.1.09)," 2020. [Online]. Available: http://www.praat.org

[15] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990, doi: https://doi.org/10.1016/0167-6393(90)90021-Z.

[16] J. B. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[17] H. Kawahara, A. d. Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on straight," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[18] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce," *p. Program available online at http://www. seas. ucla. edu/spapl/voicesauce/. UCLA*, 2009.

[19] J. Steffman and J. Cole, "An automated method for detecting F0 measurement jumps based on sample-to-sample differences," *JASA Express Letters*, vol. 2, no. 11, p. 115201, 2022.

[20] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[21] S. N. Wood, *Generalized additive models: an introduction with R*. CRC press, 2017.

[22] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2020, R package version 2.4.

[23] M. Sóskuthy, "Evaluating generalised additive mixed modelling strategies for dynamic speech analysis," *Journal of Phonetics*, vol. 84, p. 101017, 2021.

[24] P.-C. Bürkner, "brms: An R package for bayesian multilevel models using stan," *Journal of statistical software*, vol. 80, pp. 1–28, 2017.

[25] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2021, R package version 1.7.1-1. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[26] D. Makowski, M. S. Ben-Shachar, and D. Lüdecke, "bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework." *Journal of Open Source Software*, vol. 4, no. 40, p. 1541, 2019. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.01541