The rise and fall of American English pitch accents: Evidence from an imitation study of rising nuclear tunes

Jeremy Steffman[1], *Stefanie Shattuck-Hufnagel*[2], *Jennifer Cole*[1]

[1]Northwestern University
[2]Massachusetts Institute of Technology

jeremy.steffman@northwestern.edu, sshuf@mit.edu, jennifer.cole1@northwestern.edu

Abstract

Rising pitch movements associated with pitch accents are frequently described in terms of alignment and scaling; for example, L+H* versus H* accents vary in these parameters. We examine how 12 American English nuclear tunes, created by combining three pitch accents {H*, L+H*, L*+H}, and four edge tone sequences {H-H%, H-L%, L-H%, L-L%}, are distinguished in an imitative speech production paradigm. Bottom-up clustering analyses of unlabeled time-series f0 identify a robust distinction between trajectories that rise throughout (rise-only) and those with rising-falling movements (rise-fall). Additional clustering distinctions between tunes with different pitch accents are observed only in the rise-only cluster, and further reflect variation in holistic nuclear tune shape. For rise-fall movements, further distinctions in clustering are best defined by ending f0, corresponding to a boundary tone distinction {H%, L%}. With only 4 distinct clusters emerging from the imitated tunes, it appears that some tune distinctions are lost. Nevertheless, modeling trajectories with ToBI labels using a GAMM, and testing alignment of f0 turning points, reveals small differences between tunes in f0 scaling and alignment, distinguishing ToBI labels that were grouped together in clustering. We discuss these results in terms of the hierarchy of distinctions they imply, and categories of tune shapes.

Index Terms: intonation, nuclear tunes, alignment, scaling, clustering, imitative speech production.

1. Introduction

Phonological categories in AM models are often described in terms of alignment and scaling. For example, distinctions between pitch accents such as L*+H and L+H* are identified based on the alignment of an f0 event, e.g., a rise, with respect to a metrically prominent syllable. Scaling is another parameter which has been described as differentiating categories, particularly in the case of pitch accents, for example the same tonal element may be realized with a higher f0 value in one context than in another (as in H* versus L+H*). A common and recurring question in the intonation literature is if, and how, continuous parameters such as alignment and scaling of f0 map to phonological categories [e.g., 1-5]. Put differently, how much does variation in alignment and scaling reflect the implementation of discrete, categorical representations?

More generally, it is an open question whether the full set of pitch distinctions predicted by the phonological inventory of tones in the AM model of American English [6,7] are readily available for speakers to produce [8]. We address this question in an experiment testing how speakers produce distinctions among a set of tunes which are predicted to vary in alignment and scaling, testing three pitch accents (H*, L+H* and L*+H), combined with all boundary tone sequences (H-H%, H-L%, L-H% and L-L%): twelve total nuclear tunes. We examine which f0 parameters best distinguish the nuclear tune shapes produced in imitative speech, and ask if the observed distinctions map straightforwardly onto ToBI labels. The f0 trajectories are analyzed using a clustering analysis on unlabeled data. We further use GAMM modeling to examine what differences in alignment and scaling are detectable when the analysis takes pre-defined tune categories into account.

2. Methods

We adopted an imitative speech production paradigm modeled on [8]. Speakers were asked to reproduce the tunes of heard model utterances, for which f0 trajectories have been resynthesized based on [9,10]. Within a trial, the participant listens to model utterances, and then reproduces the *exposure* tune from the model utterances on a new sentence (the *imitation*), which is shown orthographically on the computer monitor. Participants are instructed to do so in a way that sounds natural to them. The model utterances and imitative utterances were syntactically similar, and all ended in a tri-syllabic, stress-initial name, on which the nuclear tune was instantiated. Model utterances were produced by two model speakers (one male, one female) and contained two sentences ("Her name is Marilyn"/ "He answered Jeremy"). The new sentences that speakers were prompted to say aloud were "She remained with Madelyn"/ "He modeled Harmony"/ "They honored Melanie".

In a trial, a participant heard two model utterances, instantiating the same exposure tune on the final tri-syllabic, stress-initial name (e.g., "Marilyn"). In each trial, the model utterances comprised one production from each model speaker, and both model sentences, for a total of four possible (2 x 2) stimulus combinations, which appeared with equal frequency throughout the experiment. Each stimulus combination was paired with each of 12 tunes (described below) for a total of 48 unique trials, repeated three times and presented in a fully randomized order for a total of 144 trials in the experiment. The experiment was completed remotely, with participants listening to stimuli over headphones/earbuds, and recording their responses with their own built-in/external microphone. 70 speakers participated in the experiment, recruited from Prolific and the Northwestern Linguistics Subject Pool, with each self-identifying as a native American English speaker (36 female, 31 male, 3 gender non-binary; mean age =22).

2.1. Materials and measurement

Stimuli were created by resynthesizing naturally produced utterances, with f0 trajectories created on the basis of straight-line approximations as described in [6,11], using 6 f0 target heights located at the same proportional location in each speaker's pitch range. Alignment of f0 movements corresponded to segmental landmarks, as shown in Figure 1A.

Utterances were segmented via text grid to identify the region predicted to carry the nuclear tune (e.g., "Melanie" in "They honored Melanie"), and the portion of the sentence preceding the nuclear tune. Text grids were force-aligned using the Montreal Forced Aligner [12], and subsequently manually checked and hand corrected when necessary. Aligned files were submitted for f0 measurement using STRAIGHT, as implemented in VoiceSauce [13,14]. Files containing likely f0 tracking errors were detected by an algorithm that computed sample-to-sample changes in f0 implemented as in [15], and flagged as likely errors those changes which exceeded f0 rate of change thresholds described in [16]. Flagged files were subsequently manually inspected, and excluded if an f0-tracking error was confirmed. In total we excluded 9% of the files on this basis (note that non-modal phonation was common and led to inaccurate sudden jumps in estimated f0). We additionally excluded two speakers (from an original total of 72), for whom poor audio quality made reliable f0 extraction difficult. We time-normalized the f0 measurements, taking 30 equidistant samples per nuclear word. We further converted f0 measures from Hz to ERB and scaled and centered each speaker's measures, effectively normalizing for differences in pitch height and pitch range. Figure 1B plots the by-speaker average time-normalized f0 trajectories for imitative productions of each exposure tune.

2.2. Analyses

Here we report on three analyses, each providing a different assessment of the distinctions present in the imitative data. First, we present the results of a clustering analysis, implementing *k*-means clustering for longitudinal data [17] (Section 3.1). Unlabeled f0 trajectories are partitioned into clusters which are iteratively optimized via cluster centroids. We selected the optimal partition of the data using the Calinski-Harabatz criterion [18], which selects as optimal the solution with the highest ratio of between to within cluster dispersion, computed over time series vectors. We tested two through ten clusters as possible partitions. Here we are effectively asking what *number of clusters* best characterizes the unlabeled data, a "bottom up" approach to discovering distinctions among imitated tunes. The analysis was carried out on speaker mean trajectories for each tune (12 trajectories per speaker).

We also assessed differences between trajectories which were labeled by exposure tune, a "top down" approach to describing contour differences, carried out with individual trial-level productions (not speaker means). First, we modeled time-normalized scaled ERB for imitations of each tune using a GAMM (Section 3.2), fit using [19,20] and predicting f0 by tune, with random effects specified using reference/difference factor smooths, comparable to random intercepts for speaker and by-speaker random slopes for tune, implemented as in [21]. Our second "top down" analysis modeled f0 turning points, no longer in normalized time, but instead in terms of (raw) temporal distance from the end of the first syllable in the nuclear word (this boundary was manually checked in the auditing of the text grids). We modeled the timing of f0 turning points (Section 3.3) using a mixed-effects regression model implemented in the Bayesian framework [22], which predicted turning point timing by exposure tune, with random intercepts for participant, and by-participant slopes for tune.

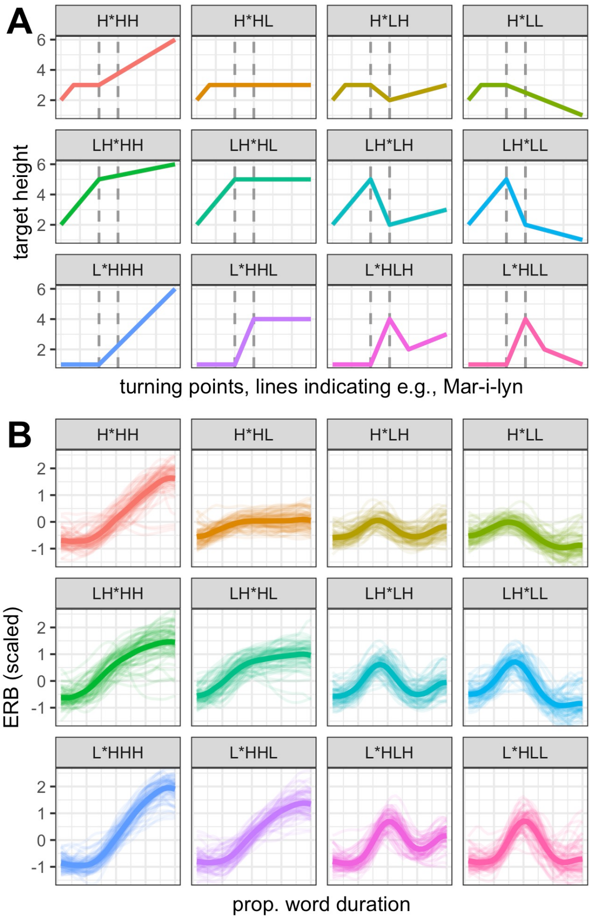

Figure 1: Schematic model tunes, with vertical lines indicating segmental landmarks (A), and mean time-normalized speaker productions (B), with thin lines indicating speaker means and the thick line indicating the grand mean. H*HH refers to H*H-H%, and so on.

3. Results

3.1. Clustering analysis

The results of the clustering analysis are shown in Figure 2. As shown in Figure 2A, the optimal partition of the data was into only two clusters, here labeled 1 and 2. Cluster 1 mainly includes imitations of tunes with a rise-fall contour, while cluster 2 mostly includes imitations of tunes which contain no fall. A second-pass clustering analysis was carried out separately for the imitations in each of the two first-pass clusters, as shown in Figure 2B with subclusters (1a, 1b, 2a, 2b). Here we examine how each cluster varies in shape, and how exposure tunes, defined based on ToBI labels, map to clusters.

First, consider clusters 1a and 1b and their mean trajectories in Figure 2B. These mean cluster shapes are best distinguished by the scaling of f0 after the initial peak f0 associated with the pitch accent: In cluster 1b, f0 falls from the initial peak to a low value, while in cluster 1a there is a much smaller f0 fall after the peak, ending in a mid-level f0. The mapping of exposure

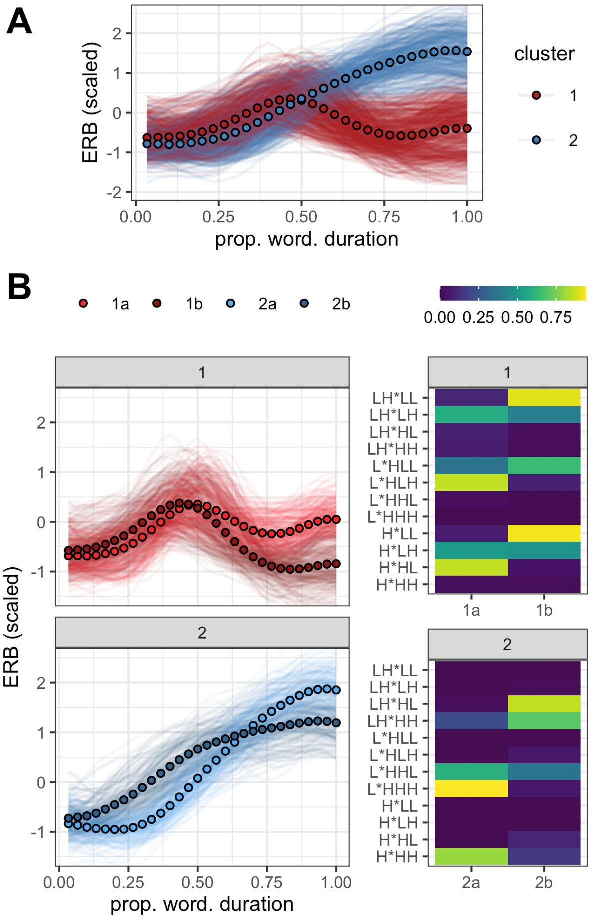

Figure 2: Cluster partitions for the full data set (A) and each sub-cluster (B). Faint lines are speaker means for each tune; dotted lines are cluster means. Heat maps at right show cluster composition in terms of the proportion of contributing tunes, indexed by color.

tunes to clusters is largely, though not entirely, based on boundary tones: imitations of LH-ending tunes tend to fall into cluster 1a, while those of LL-ending tunes fall into cluster 1b. Imitations of H*LH are evenly distributed across the two clusters. The mean trajectories of the two clusters differ slightly in alignment of the initial pitch movement (associated with the pitch accent), but more so in the scaling of the final portion of the tune, with higher (1a) or lower (1b) f0.

The f0 trajectories in clusters 2a and 2b rise throughout, and differ in two parameters: whether the rise is scooped in shape (2a) or domed (2b), and the scaling of the ending f0 in each, with cluster 2a ending higher. This distinction in shape maps onto exposure tune labels, with imitations of L*HHH, L*HHL and H*HH mostly making up cluster 2a, and imitations of LH*HH and LH*HL mostly making up cluster 2b. This distinction in scooped vs. domed rise shape has been documented elsewhere, where it is described in terms of the Tonal Center of Gravity [4,23,24] of an f0 event. This may be the basis of the distinction among the high-rising nuclear tunes in our study as well, distinguishing clusters 2a and 2b, though this merits further investigation.

In summary, the four clusters which emerge from our analysis define a hierarchy of distinctions, with rising vs. rising-falling at the highest level. A further distinction in the rising-falling subset corresponds primarily to the boundary tone sequence (cluster 1a/1b), manifest in final f0 height. For the rising subset, a distinction in rise shape (domed vs. scooped) corresponds fairly well to a distinction between the two bitonal pitch accent categories in the exposure tunes, though differences in f0 contour shape are holistic, spanning the tune.

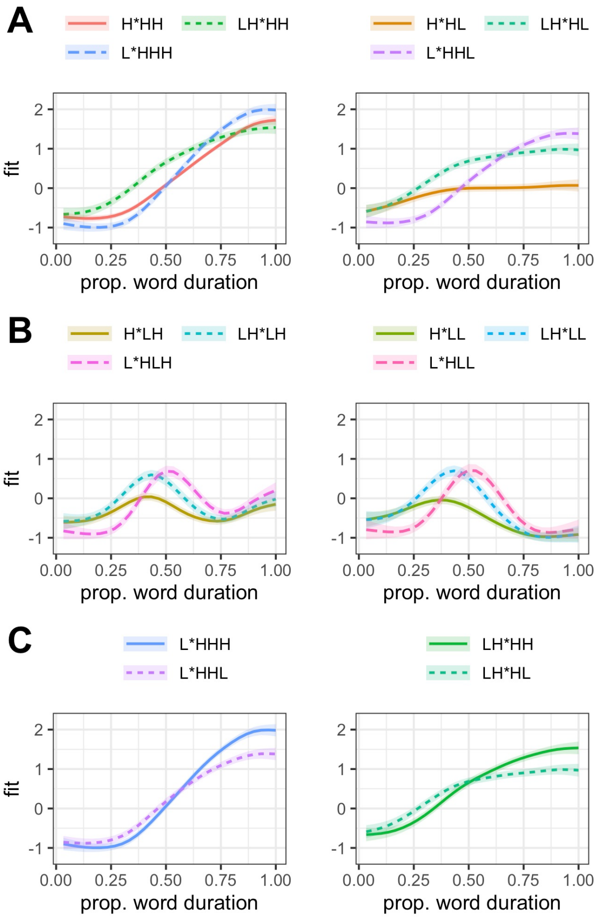

Figure 3: GAMM fits for tunes over normalized time, with 95%CI. Non-overlapping CI can be taken to indicate a significant difference across tunes.

3.2. GAMM modeling of trajectories

Given the results of the clustering analysis, we have evidence that some distinctions between exposure tunes are not well preserved by speakers in the experiment. For example, imitations of LH*HH and LH*HL mostly cluster together in cluster 2b, suggesting that they are produced without clear distinctions in shape. To further assess distinctions among tunes, we visualized GAMM smooth fits and confidence intervals (CIs), shown in Figure 3. For any pair of tunes, we take non-overlapping CIs as evidence of a reliable difference between exposure tune shapes. Figure 3A and 3B show GAMM fits for each pitch accent, grouped by edge tone sequence. We note here that each set of tunes (within a panel) shows regions that are differentiated by the GAMM analysis- and moreover each contour shows the expected distinction in shape based on the exposure tunes (see Fig. 1), save for H*HH, which has been reproduced as a scooped low-to-high rise. With respect to panel B in particular, we see that, although imitations of tunes with different pitch accents were grouped in the same

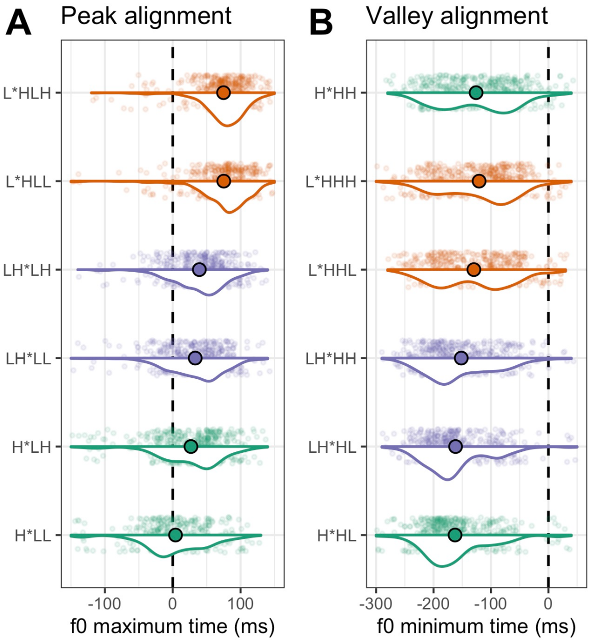

Figure 4. Alignment of f0 peaks (A) and valleys (B), with tune at left. Violin plots show the distribution, the large points show the mean. Tunes are ordered top to bottom from latest to earliest alignment. Color indicates pitch accent. Time 0 marks the end of the first syllable.

cluster (1a/1b), there are detectable differences in the region of the pitch accent in f0 scaling (H*LL vs. LH*LL) and in alignment (LH*LL vs. L*HLL). We can make analogous comparisons for the imitations of tunes that clustered together in clusters 2a and 2b, shown in panel C of Figure 3. These clusters grouped imitations of tunes with different edge tones (HL and HH); however GAMM fits show differences between these two edge tone sequences corresponding to distinctions in rise shape between scooped and domed (particularly for LH*HH/LH*HL), and in scaling of final f0 height.

3.3. Alignment of f0 peaks and valleys

Some of the differences among nuclear tunes relate to the alignment of f0 peaks and valleys. In our set of resynthesized exposure tunes, differences in f0 alignment are important for distinguishing the 3 pitch accent categories. We find marginal evidence for alignment distinctions among imitated productions in the clustering analysis, where 2 (but not 3) pitch accents are distinguished, only in the rising subset (Fig. 2B, cluster 2a/2b). We took a closer look at alignment through the timing of two f0 events with respect to the boundary between the first and second syllable in the word carrying the nuclear tune. We measured alignment of the f0 peak associated with the pitch accent rise for imitations of exposure tunes with a rise-fall shape (those with an L- phrase accent), in terms of the (often positive) distance from the syllable boundary to the f0 peak of the rise. For exposure tunes with an overall rising shape (those with H-phrase accents), we measured alignment of the valley of the rise (i.e., the rise onset) as the (often negative) distance of the f0 minimum to the syllable boundary (Fig 4). We report estimates from the statistical model which were found to have a reliable effect on alignment, assessed by inspecting the 95% posterior distribution for a given effect. When the posterior excludes 0, this indicates a clear directionality for the effect. We also report the percentage of the posterior with a given directionality [25] as "pd" (for "probability of direction"); pd = 99% indicates that 99% of the distribution has a given sign: strong evidence for an effect. With L+H* as the reference level in the peak alignment model, H* is earlier (β = -17; CI = [-27,-9]; pd = 100%) and L*+H is later (β = 42; CI = [33,50]; pd = 100%), without strong evidence for an effect of edge tone sequence (pd = 83%). In the valley alignment model, with L+H* as the reference level, L*+H shows later alignment (β = 20; CI = [10,31]; pd = 100%), with a further credible effect of boundary tone where HL boundary tones lead to earlier valley alignment (β = -11; CI = [-20,-3]; pd = 100%). There is also an interaction (pd = 100%), indicating the boundary tone effect is larger for H* pitch accents - this seems related to the fact that H*HH is imitated with a different shape in the initial portion compared to H*HL (see Fig. 1/Fig. 3); H*HH shows the latest mean alignment while H*HL shows the earliest. Especially for L+H* and L*+H, the joint influence of edge tones and pitch accent on alignment of the valley (rise onset) is a departure from what we see in peak alignment, suggesting that rise onset timing is more holistically determined by these two parameters. In both peak and valley alignment, but especially valley alignment, distributions are heavily overlapping, and effects are small in magnitude. Nevertheless, these results suggest fine-grained distinctions that were not captured in the clustering analysis.

4. Discussion

Our results suggest a hierarchy of distinctions among nuclear tunes, as reflected in imitated productions. Clustering of unlabeled f0 trajectories of imitations shows a primary partition into rise and rise-fall shapes, with a secondary partition defined by scaling of final f0. For imitations in the rise class only, this secondary partition also marks a distinction between scooped vs. domed rises, corresponding to a Tonal Center of Gravity distinction. Within these four clusters, finer distinctions in f0 alignment and scaling emerge when imitated productions are grouped by exposure tune category, some of which correspond to predicted distinctions between tonal categories (e.g., pitch accent distinctions, Fig. 3B), while others are best described in terms of holistic tune shape (domed vs. scooped rises) that integrate pitch accent and edge tone features. That finer distinctions are not captured in the clustering analysis reflects their smaller scale and variable implementation. We note here that with the present data we are not able to localize the source of noisy imitations to perception (of model tunes) or to speakers' production systems reducing or eliminating perceived distinctions. In sum, the present results show the importance of considering distinctions among tunes as part of a whole system. To evaluate the AM model of American English nuclear tunes it is necessary to consider how tunes in the inventory are implemented in relation to one another, and the extent to which they are reliably distinguished from *all other members* of the inventory. This approach gives further insight into the ways in which speakers actually make tunes distinct, some of which are predicted by the AM model, and some of which are not (e.g., domed vs. scooped rises). Further tests of tune categories in this vein will benefit from examining other acoustic correlates of intonation (duration, intensity, voice quality), as well as perceptual data, and also from putting tunes in discourse contexts, which may support or enhance certain distinctions.

5. Acknowledgements

We thank Chun Chan for technical help. This project was supported by NSF BCS-1944773.

6. References

[1] Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2010). Turning points, tonal targets, and the English L- phrase accent. *Language and Cognitive Processes*, *25*(7), 982–1023.

[2] Dilley, L. C., & Heffner, C. C. (2013). The role of f0 alignment in distinguishing intonation categories: evidence from American English. *Journal of Speech Sciences*, *3*(1), 3–67.

[3] Ladd D.R., Faulkner D., Faulkner H., & Schepman A. (1999). Constant "segmental anchoring" of f0 movements under changes in speech rate. *Journal of the Acoustical Society of America* 106(3):1543-54. 12.

[4] Niebuhr, O., d'Imperio, M., Fivela, B. G., & Cangemi, F. (2011). Are There "Shapers" and "Aligners"? Individual Differences in Signaling Pitch Accent Category. In *Proceedings of ICPhS* (pp. 120-123).

[5] Grice, M., Ritter, S., Niemann, H., & Roettger, T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics*, *64*, 90-107.

[6] Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. Doctoral dissertation, Massachusetts Institute of Technology.

[7] Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology*, *3*, 255-309.

[8] Chodroff, E., & Cole, J. (2019). Testing the distinctiveness of intonational tunes: Evidence from imitative productions in American English. In *Proceedings of INTERSPEECH* (pp. 1966-1970). International Speech Communication Association.

[9] Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5-6), 453-467.

[10] Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.48 retrieved 4 June 2021 from http://www.praat.org/

[11] Veilleux, N., Shattuck-Hufnagel, S., & Brugos, A. *6.911 Transcribing Prosodic Structure of Spoken Utterances with ToBI.* January IAP 2006. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu.

[12] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of INTERSPEECH* (pp. 498-502).

[13] Kawahara, H., Cheveigné, A. D., Banno, H., Takahashi, T., & Irino, T. (2005). Nearly defect-free f0 trajectory extraction for expressive speech modifications based on STRAIGHT. In *Ninth European Conference on Speech Communication and Technology*.

[14] Shue, Y.-L. (2010). *The voice source in speech production: Data, analysis and models.* Doctoral Dissertation, University of California, Los Angeles.

[15] Steffman, J. (2021). f0 jump detector (Version 1.0.1) [Computer software]. https://github.com/jsteffman/f0-jumps/tree/v1.0.1

[16] Sundberg, J. (1973). Data on maximum speed of pitch changes. *Speech transmission laboratory quarterly progress and status report*, *4*, 39-47.

[17] Genolini, C. Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, 65(4), 1-34.

[18] Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1-27.

[19] Wood, S.N. (2017) Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC.

[20] van Rij J., Wieling M., Baayen R., van Rijn H. (2020). "itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs." R package version 2.4.

[21] Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, *84*, 101017.

[22] Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

[23] Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology*, *3*(2), 337-383.

[24] Barnes, J., Brugos, A., Veilleux, N., & Shattuck-Hufnagel, S. (2021). On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity. *Journal of Phonetics*, *85*, 101020.

[25] Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, *4*(40), 1541.