The primacy of the rising/non-rising dichotomy in American English intonational tunes

Jennifer Cole, Jeremy Steffman

Northwestern University
jennifer.cole1@northwestern.edu, jeremy.steffman@northwestern.edu

Abstract

In American English, phrase-final pitch trajectories have been described as resulting from a sequence of three tonal elements whose combinations define an inventory of phonologically contrastive *nuclear tunes* [1]. We investigate the distinctive status of nuclear tunes, testing imitative production of sentences paired with one of 8 nuclear tunes, and testing pairwise perceptual discrimination of the same tunes. Results from group- and individual-level clustering analyses of F0 trajectories of imitated tunes reveal maximally 5 distinct tunes, with the most robust distinctions between two tune classes: rising and non-rising. Converging results are obtained from perceptual discrimination. A further finding is that the phonetic distance between tunes is a good predictor of discrimination accuracy, but accuracy is better than predicted for pairwise discrimination across the rising/non-rising classes, and worse than predicted for tunes grouped together in the rising class. These results suggest a robustness hierarchy of tune distinctions with a primary rising/non-rising distinction. This hierarchy reflects holistic shape distinctions, but does not align with the proposed tripartite composition of tunes.

Index Terms: intonational phonology, nuclear tunes, imitation, intonation production, intonation perception

1. Introduction

In American English (AE) intonation, the pitch pattern in the final region of a prosodic phrase conveys pragmatic meaning. In the phonological analysis proposed in [1], this pitch pattern is defined by the concatenation of three components: the pitch accent, phrase accent, and boundary tone, specified in terms of the tonal primitives H(igh) and L(ow), which determine relative pitch targets for the associated syllables. Interpolative pitch movements between successive tonal targets yields dynamic pitch patterns that extend from the location of the nuclear stressed syllable to the end of the phrase.

Setting aside the downstepped high tone (!H) and bitonal pitch accents (L+H*, L*+H, H+!H*) in the feature inventory proposed for AE intonation [2], this system generates a set of 8 tonally distinct tunes, mapping onto 8 distinct pitch contours that are in principle available for encoding pragmatic meaning contrasts. We refer to this as the set of "basic" tunes. This paper is concerned with the phonological status of the tunes in this basic set, i.e., as representing distinct categories in mental representation for which there are systematically distinct acoustic realizations. Some evidence for distinctions among discrete tune categories is found in studies on intonational meaning that show associations between tune and meaning, yet although there are many such studies in the literature, none address the system of contrast among more than a few tunes and many do not specify the phonological tones or characteristic F0 contours of the tunes they analyze. Thus, even though there is some support for the general claim that certain tunes are associated with certain distinctions in pragmatic meaning [e.g., 3, 4, 5], there is not yet solid empirical support for the claim of an 8-way phonological contrast in the basic tune inventory. Ultimately, what is needed is empirical evidence for the associations between tunes and meaning establishing the relationships of contrast, yet this is not a straightforward undertaking in the analysis of intonation due to two critical questions for which research has yet to reach consensus [5, 6]: (1) What are the pragmatic meaning distinctions conveyed by intonation? (2) What are the set of tunes that can be produced and perceived as distinct from one another, and which are therefore available for encoding pragmatic meaning?

This study addresses the second question, investigating distinctions in the production and perception of the 8 basic tunes of AE. For production evidence, we use imitative speech rather than an alternative approach using discourse prompts that set up pragmatic contexts to elicit distinct tunes, due to the lack of a model of pragmatics covering all 8 basic tunes, and to avoid uncertainty about how participants might interpret pragmatic meaning from a discourse prompt. Our production experiment involves presenting participants with auditory models of each tune, then asking them to reproduce the heard tune on a new sentence. The participant must abstract a pitch melody from the model utterance and apply it in the production of a new sentence. We assume that this process involves assigning a linguistic representation to the heard melody, which will be reflected in the acoustic properties of the imitated melody. For evidence of an 8-way distinction in perception, we use an AX discrimination task, examining listeners' same/different responses to tune pairs using the auditory model utterances from the imitative production experiment.

Our hypothesis for both experiments is that if the basic tune inventory is part of a participant's knowledge of AE intonation, they will be able to perceive each tune as distinct from the others, and their imitated productions will reveal acoustic distinctions among all 8 tunes. Conversely, failure to perceive and/or produce a systematic acoustic distinction between any two or more tunes would suggest marginal or missing contrasts in the hypothesized tune inventory. To test this hypothesis in production, we submit imitated f0 trajectories to group- and individual-level clustering analyses to determine the number of robust distinctions among them. To test the status of an 8-way distinction in perception, we analyze same/different responses from the AX discrimination task in relation to the tune labels of the paired stimuli. As discussed below, findings from both experiments point to a hierarchy of tune distinctions, with a robust distinction between rising and non-rising tunes, and less robust distinctions within each of those classes.

2. Methods

Speech production experiment. We elicited imitative productions of nuclear tunes using the experimental paradigm from [7]. The set of 8 nuclear tunes are formed over combinations of a monotonal pitch accent, phrase accent and boundary tone (tunes abbreviated as HHH, HHL, HLH, HLL, LHH, LHL, LLH, LLL). Participants heard model utterances with resynthesized f0 trajectories representing the 8 nuclear tunes, imitated the heard tune on each trial, reproducing it in a new sentence presented in text format on the computer screen. Participants were encouraged to reproduce the tunes in a way that sounded natural to them. The sentences in the model utterances and the new sentences were syntactically similar and ended in a trisyllabic, stress-initial name on which the nuclear tune was instantiated. Model utterances were produced by 2 speakers (one male, one female) in 3 sentences ("Her name is Marilyn"/ "He answered Jeremy"/ "He quoted Helena"). The new sentences that participants produced were "She remained with Madelyn"/ "He modeled Harmony"/ "They honored Melanie".

On each trial, participants heard 3 model utterances instantiating the same nuclear tune. F0 was resynthesized for natural productions of the model utterances, using PSOLA in Praat [8,9] with a linear f0 decline over the preamble and implementing straight-line approximations of the nuclear tunes, as shown schematically in Figure 1. The resynthesized f0 contours differed from those used in [7] in both f0 scaling (lower peak f0, most notably for HHH, LHH) and in the alignment of f0 turning points to consistent segmental landmarks rather than at fixed temporal intervals. Tunes were implemented with five target f0 values, located in each model speaker's pitch range. The scaling and alignment of resynthesized tunes were based on examples from [1, pp. 391-401] and online training materials [3] and were additionally judged to sound appropriate for each tune by two expert listeners trained in intonation annotation using the ToBI system (including the first author).

Participants were 30 self-reported native speakers of American English (18 female, 11 male, 1 gender non-binary, mean age = 21), recruited from the Northwestern University subject pool (22) and from Prolific (8). They participated remotely, using their own computer, microphone, and headphones/earbuds. There were 144 trials (8 tunes × 18 trials per tune). The 18 trials for a given tune differed in the order of the 3 model sentences (6 orders, balanced for model speaker gender), and in the target sentence (3 sentences). F0 in the participants' imitative productions was measured using STRAIGHT in Voicesauce [10,11]. Textgrids were force-aligned [12], individually inspected, and manually corrected where needed. F0 estimates were taken from the nuclear accented word, and in the preceding (preamble) portion of the sentence. A hybrid automated/manual f0 error detection procedure resulted in the exclusion of 11% of the utterances, for a total of 3,798 imitative utterances analyzed (f0 samples were flagged as an error when exceeding f0 rate-of-change thresholds from [13] – non modal phonation was a frequent source of errors).

We performed two clustering analyses on the time-series F0 measures. Both were implemented using *k*-means clustering for longitudinal data [14]. 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 [15], 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 of the data. In this analysis we are effectively asking what *number of clusters* best characterizes the unlabeled data, a "bottom up" approach to discovering distinctions among imitated tunes.

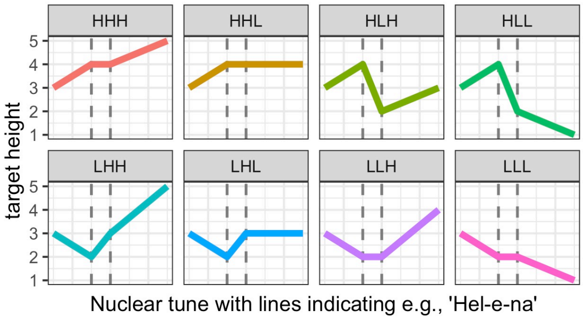

Figure 1: Schema for the models tunes. HHH indicates H*H-H% and so on. Alignment with syllable boundaries in indicated by the dashed vertical lines.

The first clustering analysis was carried out on participant mean trajectories for each tune (a single mean trajectory for each of the 8 model tunes, from each participant), allowing us to characterize the data set as a whole, with an equal contribution from each participant. Our second analysis is focused on inter-speaker variation, clustering on the imitated trajectory on each trial, separately for each participant. This analysis thus asks how an individual partitions their productions into clusters, and how many clusters best characterizes their data. In reporting the individual clustering results, we focus on how participants vary and commonalities in their clustering solutions.

Speech perception experiment. The perceptual salience of the input tunes was tested by listeners in an AX discrimination task, using model utterances from the speech production experiment (8 tunes produced by 2 speakers, on 3 different sentences, for 48 unique stimuli). 30 different native speakers of American English, recruited on Prolific, participated remotely (14 female, 15 male, 1 gender non-binary, mean age = 23). On each trial, participants were presented with recordings of a tune pair and asked to respond, by mouse click on a labeled button, if the two tunes were the same or different. The inter-stimulus-interval was 500 ms. Participants were instructed to focus on the intonational melody of the utterance. Tunes were paired with each other in all possible order-sensitive combinations yielding 64 tune pairs (8 x 8 tunes). This 64-tune list was repeated, for 128 trials in total. For both tunes in a given trial, the model speaker voice and the model sentence were the same. Model speaker and sentence varied across trials and were combined with tune pair in three counter-balanced lists. Ten of the 30 total participants were randomly assigned to each list, hearing different model speakers and sentences across randomized trials. All possible combinations of model speaker, sentence and tune were attested across the three counterbalanced lists.

We analyzed responses to order-insensitive tune pairs (e.g., combining responses to HHH-HHL & HHL-HHH) to assess how accurately listeners discriminated tune pairs. Bayesian logistic regression in Stan [16] was conducted to model variation in listeners' responses ("same" or "different"), as a function of tune pair, with random intercepts for listener, and weakly informative normal priors for both the intercept and

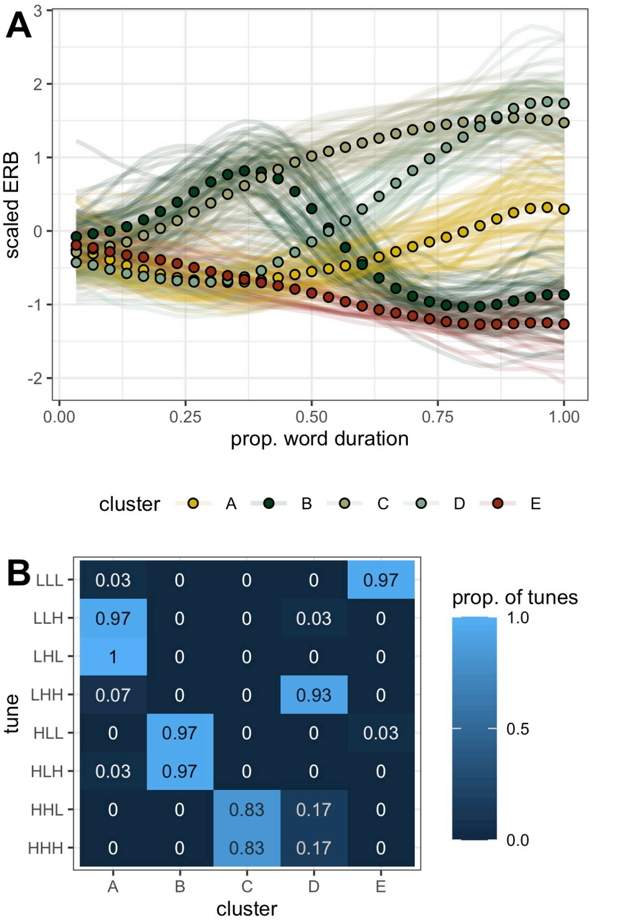

Figure 2, Panel A: the clustering solution, with cluster means shown by the dotted lines and contributing trajectories as lighter lines. Panel B: the mapping from tune to cluster with tunes in rows, clusters in columns and the proportion of each tune in each cluster indicated by the color scale and number in a cell.

fixed effects. Results are reported here only for "different" trials, as performance on same-tune trials was near ceiling for all tune pairs.

3. Results

Group-level clustering. The optimal solution for the clustering algorithm is one with 5 clusters, shown in Figure 2A. We assess the composition of these clusters in terms of the mapping from imitated tunes to clusters, associating each imitation with the label of the model tune it imitates, and identifying the cluster it was assigned to in the optimal clustering solution. The mapping of imitated tune to cluster is shown in the heat map in Figure 2B, with cluster A composed primarily of LLH and LHL tunes (97% and 100% of those tunes respectively), indicating poor differentiation of these two categories in their imitated productions. Cluster B is similarly composed of HLL and HLH tunes, which are likewise not well differentiated in production. The same can be said for HHH and HHL which make up cluster C. In comparison to clusters A-C, cluster D consists primarily of LHH tunes, with minimal contributions from imitations of HHH and HHL, and Cluster consists of mostly LLL tunes. To summarize, the group level clustering results show two tunes, LLL and LHH, as robustly distinguished in imitations, while among the other 6 tunes in the inventory only 3 distinctions emerge (Clusters A, B, C), each effectively merging a pairwise distinction among the model tunes presented as stimuli.

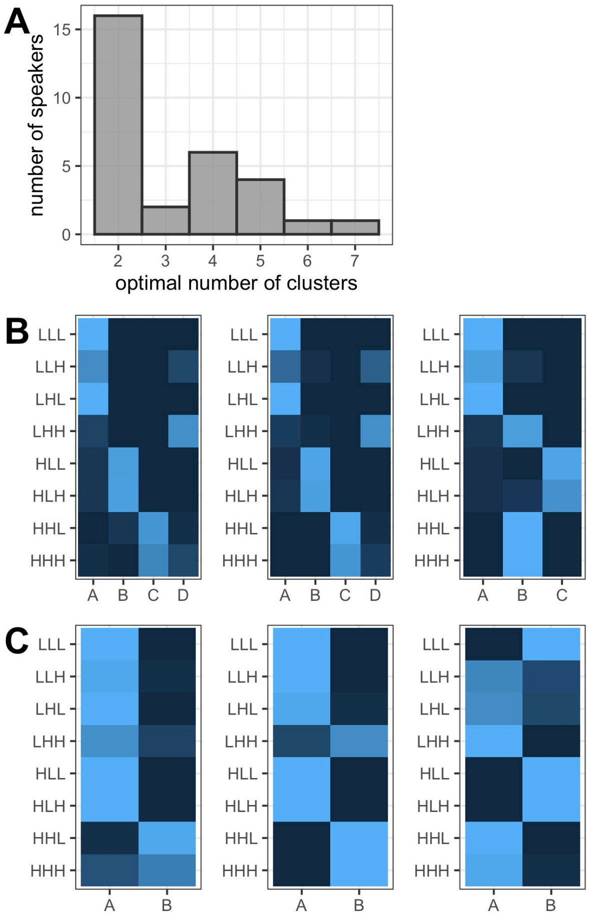

Figure 3, Panel A: the distribution of clustering solutions for individual speakers. Panel B: examples of individuals with 4 or 3 cluster solutions. Panel C: examples of individuals with 2 cluster solutions. The color scale in heat maps is the same as in Figure 2.

Individual clustering. Individual-level clustering results reveal substantial variation in terms of the optimal clustering solution. The distribution of these solutions by speaker is presented in Figure 3A, and qualitatively assessed in comparison of the tune-to-cluster mapping in the aggregated data (Figure 2B).

As is clear in Figure 3A, though some speakers evidenced a five-cluster solution (for which the tune-to-cluster mapping was highly similar to that in Figure 2B), many speakers differentiated fewer than five tunes in their imitations. Heat maps for representative speakers with four- and three-cluster solutions are shown in Figure 3B. The four-cluster speaker at left groups LLL with the two low-to-mid rising tunes LHL and LLH that combine in Cluster A in group-level analysis. The middle panel of Figure 3B evidences a similar pattern, with LLH imitations additionally split between this low-to-mid rising cluster and the cluster made up of LHH. The three-cluster speaker at right additionally merges LHH, HHH, and HHL, all tunes that rise to a high f0 target, into a single cluster.

Two-cluster speakers, the most common solution in the data, evidence various partitions of the tunes into two clusters,

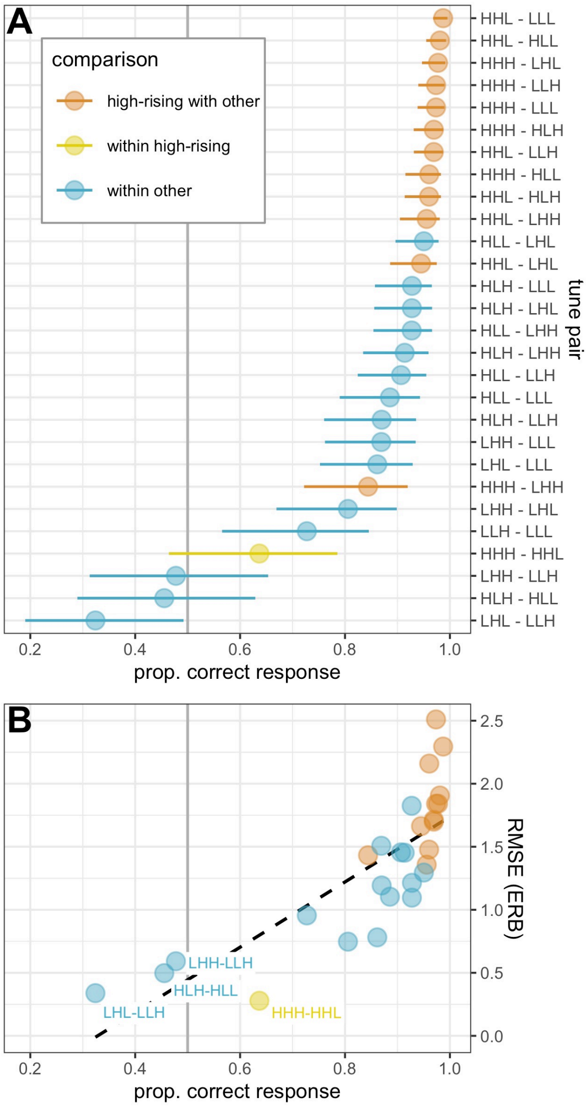

Figure 4, Panel A: Model fit for accuracy in the perception experiment, plotting correct ("different") responses on the x axis, the dashed line indicating chance. All tune pairings are shown on the y axis, sorted by accuracy. Error bars show 95% CrI. Panel B: accuracy plotted against RMSE, with the four lowest accuracy tune pairs labeled.

with the commonality that these partitions separate tunes primarily by the extent to which they rise over the course of the nuclear region. The leftmost panel in Figure 3C clusters mainly HHL and HHH, the "high rising" tunes, into a single cluster, distinct from imitations of all other tunes, which are not distinguished and cluster together. Other speakers have an expanded "rising" cluster that includes LHH (Fig. 3C, middle) and even the low-to-mid rising tunes, LHL and LLH (Fig. 3C, right). Common across speakers is a distinction between rising tunes and others, with variation in the composition of the rising tune class based on rise height/shape.

Speech perception/ AX Discrimination. The speech perception results align with our production results in showing that the same tune pairs that merge in the group-level clustering solution are also among the most poorly discriminated tunes in the AX task: LHL-LLH, HLH-HLL, and HHH-HHL. These three pairs are discriminated at or below chance, based on the Bayesian model estimate of correct (same/different) response that includes 0.50 (Figure 4A).The tune pair LHH-LLH is also poorly discriminated, a surprising result in light of the group-level imitation data where LHH defines a cluster by itself (e.g., Figure 3C). Other tunes are discriminated above chance, but with a range of accuracies, partially corresponding to whether or not one of the tunes in question includes a member of the high-rising set {HHH, HHL}. When the high rising tunes are compared to others, discrimination accuracy tends to be very high (orange points in Figures 4A, B), while for pairs of tunes outside of the set {HHH, HHL}, accuracy is overall lower (blue points). Figure 4B plots discrimination accuracy for a pair of tunes against their phonetic distance, computed as root mean squared error (RMSE) in ERB, at each step in the time-normalized f0 trajectory. The regression line to the data captures the predicted relationship between ERB and perceptual discrimination. Fig. 4B also shows that tune pairs comparing high-rising to other tunes (orange) trend above the predicted line, showing higher discrimination accuracy than predicted.

4. Discussion

We sought empirical evidence for a hypothesized 8-way distinction among AE "basic" nuclear tunes. Three main findings emerge. **First**, group-level clustering of imitated tunes provides evidence for a maximum of 5 distinct tunes that differ in shape, merging predicted distinctions for three tune pairs. **Second**, clustering results for individual speakers show a near-universal pattern, with all but one speaker merging the high-rising tunes {HHH, HHL} into a single cluster. For many speakers, this cluster expands into a more generalized "rising" cluster with the inclusion of LHH, and sometimes also LHL and LLH. **Third**, the perception results show that pairwise discrimination of tunes correlates with acoustic distance. Production and perception results align showing below chance discrimination for three tune pairs that merge in the group-level clustering analysis: low-to-mid rising {LHL, LLH}, falling {HLL, HLH} and high-rising {HHH, HHL}. The low-rising pair {LHH, LLH} is also poorly discriminated, as predicted from their phonetic distance, which makes unexpected the separation of this pair in the clustering analyses. This finding suggests a special status for LHH, the prototypical rising tune used in yes/no questions. The high-rising tune pair {HHH, HHL} also stands out, with discrimination between them being worse than predicted from their phonetic distance, and discrimination of either one with other tunes being often better than predicted by phonetic distance. Together, these results suggest a hierarchy of tunes based on their distinctiveness in imitative production and perception. At the top is the high-rising tune class, with the highest final f0 values (in the stimuli and imitated productions), and which is robustly distinct from other tunes for all our participants . Low-rising tunes group with this class for some speakers, from which a generalized rising tune type emerges. Among non-rising tunes two types emerge: a Rise-Fall class {HLH, HLL} and the Low-Fall tune LLL. This hierarchy, and the overall relative distinctiveness of tunes in our data, does not align with the proposed tripartite composition of tunes, but can be described in terms of holistic pitch trajectories.

5. Acknowledgements

We thank Chun Chan, Stefanie Shattuck-Hufnagel and Lisa Cox for contributions to this work This project was supported by NSF BCS-1944773.

6. References

[1] Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. Massachusetts Institute of Technology. Reproduced by Indiana University Linguistics Club, Bloomington.

[2] Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The Original ToBi System and the Evolution of the ToBi Framework. In S.-A. Jun (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 1–37).

[3] Hirschberg, J. (2004). Pragmatics and Intonation. In L. R. Horn & G. Ward (Eds.), The Handbook of Pragmatics (pp. 515–538). Malden, MA: Wiley-Blackwell.

[4] Prieto, P. (2015). Intonational meaning. Wiley Interdisciplinary Reviews: Cognitive Science, 6(4), 371–381.

[5] Westera, M., Goodhue, D., & Gussenhoven, C. (2020). Meanings of Tones and Tunes. The Oxford Handbook of Language Prosody, (March), 442–453.

[6] Pierrehumbert, J. (2000). Tonal elements and their alignment. In M. Horne (Ed.), *Prosody: Theory and Experiment* (pp. 11–36). Kluwer Academic Publishers, The Netherlands.

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

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

[9] 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/

[10] 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*.

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

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

[14] 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.

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

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