Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Shape matters: Machine classification and listeners' perceptual discrimination of American English intonational tunes

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

[1]Northwestern University
[2]Cornell University
jennifer.cole1@northwestern.edu, tilsen@cornell.edu,
jeremy.steffman@northwestern.edu

Abstract

In Autosegmental-Metrical models of intonational phonology, pitch accents, phrase accents and boundary tones may combine freely to create a predicted set of phonologically distinct phrase-final "nuclear" tunes. In this study we ask if an 8-way distinction in nuclear tune shape in American English, predicted from combinations of 2 (monotonal) pitch accents, 2 phrase accents and 2 boundary tones, is manifest in speech production and in speech perception. F0 trajectories from an imitative speech production experiment were analyzed using (i) neural net classification, and (ii) human listeners' perceptual discrimination of the model utterances. Pairwise classification accuracy of the imitative productions is highest for tune pairs that differ in holistic shape (high-rising vs. rise-fall), and poorest for tunes with the same shape that differ in (higher vs. lower) final f0. Perception results show a similar pattern, with poor pairwise discrimination for tunes that differ primarily, but by a small degree, in final f0. Together the results suggest a hierarchy of distinctiveness among nuclear tunes, with a robust distinction based on holistic tune shape, which only partly aligns with distinctions in tonal specification, and a weak/poorly differentiated distinction between tunes with the same holistic shape but small differences in final f0.

Index Terms: intonation production, intonation perception, nuclear tunes, neural net classification, deep learning

1. Introduction

In tone sequence models of intonation such as the Autosegmental-Metrical (AM) model, F0 movements are decomposed into high (H) and low (L) tonal targets, which link to positions with phrasal prominence or domain edges, and which combine to form different phrasal f0 trajectories or "tunes". For example, the widely used AM model for American English intonation [1,2] the basis for the ToBI intonation annotation system [3], contains an inventory of 5 pitch accents used to mark phrasal prominence (monotonal H*, L*, plus 3 bitonal accents), 2 phrase accents marking the edge of an intermediate phrase (H-, L-), and 2 boundary tones marking the edge of an intonational phrase (H%, L%).[1] The sequence of the final pitch accent in the intonational phrase, referred to as the nuclear pitch accent, followed by the phrase accent and boundary tone constitute the "nuclear tune" of an utterance. Taking all possible combinations of 5 pitch accents, 2 phrase accents and 2 boundary tones, the AM model of American English predicts 20 phonologically distinct nuclear tune shapes, which may be used by speakers and listeners to signal distinct discourse meanings.

Empirical investigations of the AM model for American English have focused on distinctions among pitch accents [e.g., 5-8] and phrase accents [9], with relatively less work examining nuclear tone sequences (but see [10]) and their phonetic implementation. With the goal of addressing this gap, [11] tested the distinctions among a subset of 8 nuclear tunes in American English (those using only the monotonal pitch accents) through the analysis of imitated intonation. In that study, participants heard a set of short sentences resynthesized with one of 8 nuclear tunes (the *model* tunes), and reproduced the heard tune on a new sentence presented orthographically (the *imitations*). Distinctions among f0 trajectories of the imitations were analyzed using *k*-means clustering for time-series data, which identified five clusters, each having a distinct mean f0 trajectory. Of these five clusters, only one mapped neatly onto a distinct tune from the set of model tunes (the mid-plateau H*H-L%); the remaining four clusters comprised imitations of two or more model tunes. The five-cluster solution reflected a loss of distinction between tunes of three types: steep-rising tunes ending on a high f0 {H*H-H%, L*H-H%}, rise-fall tunes {H*L-L%, H*L-H%}, and low-rising tunes ending on a mid or mid-low f0 {L*L-H%, L*H-L%}.

The results from [11] suggest that speakers in that study were operating with fewer tunes than were hypothesized in the model tune set, making distinctions based on overall tune shape and final f0 (high/mid/low), rather than individual tone components. We raise two questions about these results. The first question concerns tune scaling. The high-rising tunes in [11] ended in very high f0, resulting in f0 excursions much larger than those of the other tunes. The very large f0 excursions of the steep-rising tunes in that study, which were based on an f0 scale obtained from natural productions, may have drawn participants' attention away from the smaller distinctions in f0 that distinguished other tunes from one another. If so, we may expect more tune distinctions to be preserved when model tunes are resynthesized to avoid large scaling differences. Our second question concerns the use of *k*-means clustering to identify groups of similar f0 trajectories in the imitated data. The optimization method used in selecting the clustering solution finds the grouping of data that maximizes the distance between clusters, while minimizing distance among items belonging to

[1] Some authors [e.g., 4] further distinguish a downstepped high tone (!H), which we set aside here.

10.21437/SpeechProsody.2022-61

the same cluster. With this algorithm, f0 trajectories with different shapes that are nonetheless close in f0 space may be grouped in the same cluster, while those with larger differences are more likely to be grouped in different clusters. Would the smaller f0 differences between "lost" tune distinctions in the clustering analysis emerge using a different method for evaluating distinctions among imitated tunes?

The present paper addresses these questions in a follow-up study, using the imitative speech production paradigm from [11], now modifying the resynthesized f0 trajectories of the 8 model tunes to reduce scaling differences among them. First, distinctions in the f0 trajectories of imitations were assessed using classification analysis, with bidirectional Long-Short-term-Memory (LSTM) neural networks trained to classify f0 trajectories of the imitations in the 8 categories of the model tune set. LSTMs are a type of recurrent neural network widely used in 'deep learning' and are especially successful at pattern recognition in sequence data, including speech and language [12, 13]. Further hierarchical clustering over the classification output provides a model of the relative distinctiveness among the 8 classes of imitated tunes. Second, we compare the classification results with results from a perceptual discrimination experiment with human listeners to evaluate the role of perceptual factors in the imitation of input tunes. Below we show that, when model tunes are scaled to a similar f0 range, machine classification of the imitations is overall very good (65% accuracy). Yet the tune pairs that the classifier most often confuses are mostly the same tune pairs that are very poorly discriminated by human listeners. Notably, two of the three tune pair distinctions that were lost in the clustering analysis of [11] are among the most poorly discriminated and least accurately classified in the present study. Based on these findings, we discuss the parameters in tune shape that are robustly distinguished, and those that are not, and consider the implications for the phonological representation of intonational tunes and their perceptual salience.

2. Methods

2.1. Speech production experiment

Imitative productions of the 8 nuclear tunes formed over combinations of a monotonal pitch accent, phrase accent and boundary tone (tunes now abbreviated as HHH, HHL, HLH, HLL, LHH, LHL, LLH, LLL) were elicited using the experimental paradigm from [11]. Speakers heard model utterances with resynthesized f0 trajectories representing the 8 tested tunes. On each trial, speakers imitated the heard tune, reproducing it in a new sentence presented 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, ending in a trisyllabic, stress-initial name on which the nuclear tune was instantiated. Model utterances were produced by 2 model speakers (one male, one female) with 3 sentences ("Her name is Marilyn"/ "He answered Jeremy"/ "He quoted Helena"). The new sentences that participants said aloud were "She remained with Madelyn"/ "He modeled Harmony"/ "They honored Melanie".

In each trial, the participant heard 3 model utterances instantiating the same nuclear tune. F0 was resynthesized for the model utterances using PSOLA in Praat [14], with a linear f0 decline over the preamble and implementing straight-line approximations of the nuclear tunes, shown schematically in

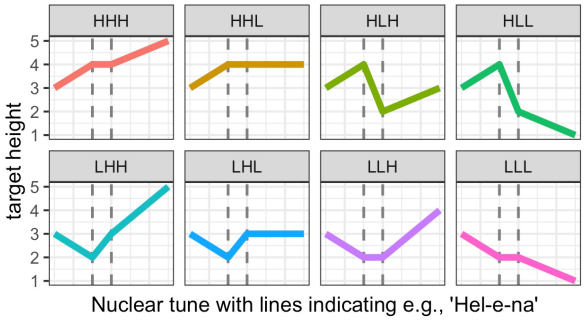

Figure 1: Schema for the models tunes

Figure 1. The resynthesized f0 contours differed from those used in [11] in both f0 scaling (lower peak f0, most notably for HHH, LHH) and in the alignment of f0 turning points to consistent segmental landmarks. Tunes were implemented using five target f0 values located in each model speaker's pitch range. The scaling and alignment of resynthesized tunes were based on examples from online training materials [3], and were judged to sound appropriate for each tune by two expert ToBI-trained listeners (including the first author).

30 self-reported native speakers of American English (18 female, 11 male, 1 gender non-binary, mean age = 21) were recruited from the Northwestern University subject pool (22) and from Prolific (8). They participated remotely in the experiment, using their own computer, microphone, and headphones/earbuds. There were 144 trials (8 tunes x 18 trials per tune). The 18 trials for a given tune differed in the order of the 3 model sentences (6 orders, balanced for gender of the model speaker), and in the target sentence (3 sentences). F0 in the participants' imitative productions was measured using STRAIGHT in Voicesauce [15, 16]. Textgrids were force-aligned [17], individually inspected, and manually corrected where needed. F0 was measured in 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 [18] – non modal phonation was a frequent source of errors).

To test whether participants produced distinct f0 patterns for all 8 input tunes, we evaluated distinctions among imitations based on the accuracy of a bidirectional LSTM neural net classifier that assigned f0 trajectories of imitations to one of 8 classes corresponding to the input tune labels. If tunes are accurately imitated, with distinct f0 trajectories reliably implemented for different tunes, classification performance should be optimal, with all imitations of a given tune assigned to the same class. Errors in the classifier output (e.g., imitations of HLL are assigned to the HLH class) are expected if imitations fail to reliably implement the distinct f0 pattern for a given tune. Frequent pairwise errors (e.g., HLL identified as HLH, and vice-versa) would reflect a loss of distinction between a pair of tunes.

Average classification accuracies for each category, along with average between-category misclassification rates, were calculated over 20 repetitions of a training-testing procedure. In each repetition, the data were randomly partitioned into training

(45%), validation (10%), and test (45%) subsets.[1] Various input representations of the f0 trajectory were tested.[2] Here we report only the combination of parameters which yielded the highest average accuracy in classification: time-normalized ERB at two time steps (*x* & *dx*), in the nuclear word only.

Agglomerative hierarchical clustering was used to infer groupings among the 8 tune classes based on average proportions of misclassfied trials, using the distance metric $\delta(A,B) = 1 - \mathrm{P}(A,B)$, where $\mathrm{P}(A,B)$ is the proportion of trials where tune A is classified as B. Tune pairs that are more often confused in the classifier output will be separated by smaller distances in the hierarchical clustering analysis. The overall hierarchical structure shows how the tune classes are dispersed in f0 space.

2.2. Speech perception experiment

The perceptual salience of the input tunes was tested by human listeners in an AX discrimination task, using model utterances from the speech production experiment. There were 8 tunes (shown in Fig. 1), produced by 2 model 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, if the two tunes were the same or different, with an inter-stimulus-interval of 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, yielding 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 3 counter-balanced lists. 10 participants were randomly assigned to each list, hearing different model speakers and sentences across randomized trials, with all possible combinations attested across the 3 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 [20] 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 fixed effects. Results are reported only for "different" trials; performance on same-tune trials was near ceiling for all tune pairs.

3. Results

Neural network classification accuracy of imitative productions is high overall, with 65% correct classification of tunes (chance = 12.5%). Similarly, perceptual discrimination for most tune pairs was well above chance (mean 80% correct; chance = 50%). However, for both the classifier and human listeners, certain tune pairs were frequently confused, as shown in the confusion matrices over tune pairs in Figure 2a. Figure 2b shows there is a relationship between the perceptual discriminability of the model tunes and classification accuracy for the imitations. In addition, confusions among imitations are less frequent than confusions among the model tunes that they were meant to imitate. This suggests that human listeners were unable to take full advantage of f0 information marking tune distinctions in the model utterances. It is also possible that in imitating the model tunes, speakers enhance acoustic distinctions between tunes, providing the classifier with extra information beyond what was available in the resynthesized model utterances.

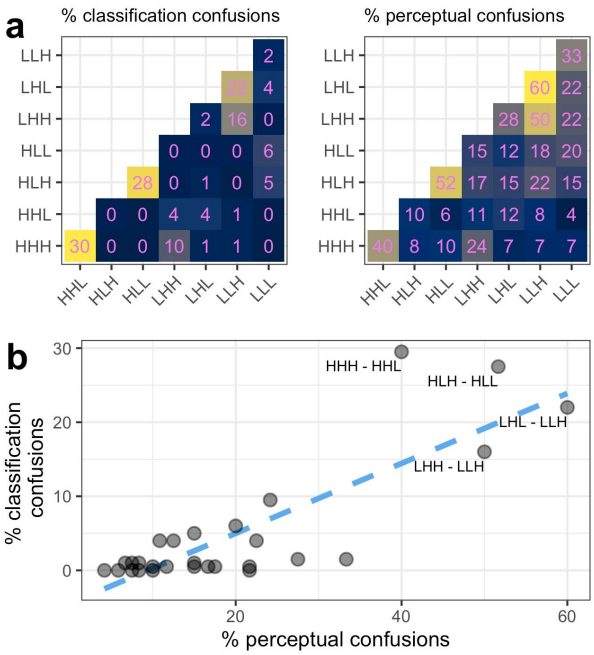

Figure 2: Heat maps showing the percentage of NN classification confusions and listener perceptual confusions for tune pairs (a), and their correlation (b).

Which tunes are confusable? From Figure 2b, four pairs of tunes stand out as being the most poorly discriminated, both in classification of the imitative production data and perception of the model tunes. These pairs are: {HHH, HHL}, {HLH, HLL}, {LHL, LLH} and {LHH, LLH}. Taking a closer look first at the imitative production data, hierarchical clustering of the classifier output (Figure 3a) shows that the tunes in these confusable pairs define clusters in the similarity space defined by the classifier output. The tunes in the high-rising pair {HHH, HLL} are the least separable (smallest distance), followed by the rise-fall {HLH, HLL} and low-to-mid rising pair {LHL, LLH}. The low-to-high rising tune LLH joins the low-to-mid

[1] The 8 tune categories were balanced within each subset. The classification networks consisted of an input layer and two bidirectional LSTM layers of 200 units, each followed by a 50% dropout layer. These were followed by a fully connected layer, a softmax layer, and a classification layer. The Adam training algorithm was used [19] with L2 regularization 0.001, learning rate 0.0001, and validation patience 20 epochs.

[2] Input representations that were tested varied in the use of (1) time-normalized vs. raw-time measurements, (2) F0 estimates in speaker-centered Hz or ERB units, or autocorrelograms (vectors of correlations between a frame of the signal with itself at all possible lags); (3) F0 estimates at each sample *x*, the difference between *x* and the following sample (*dx*), or both (*x* & *dx*), and (4) the whole utterance, just the preamble, or just the nuclear word.

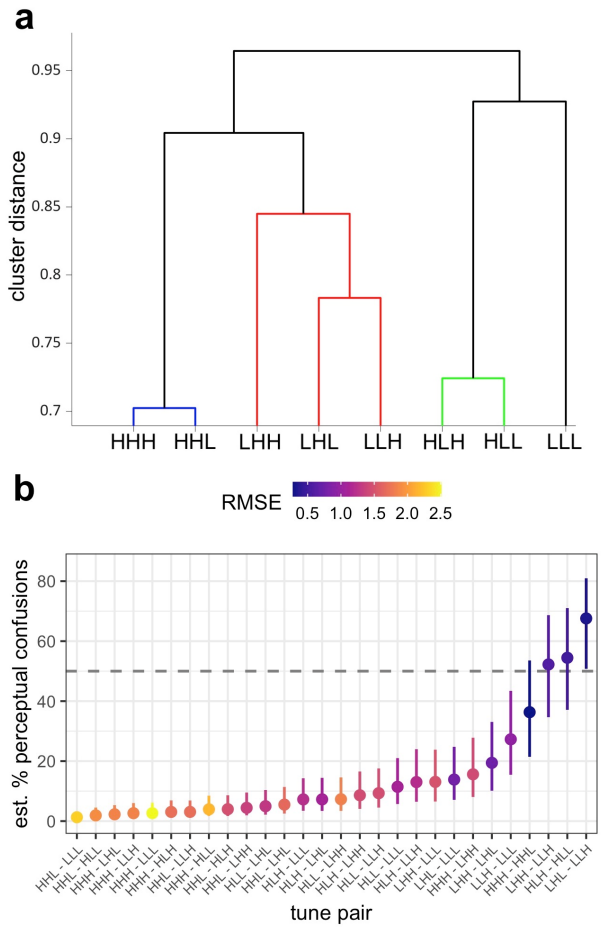

Figure 3: Hierarchical clustering for the speech production data (a), and model estimates and 95% credible intervals for tune-pair discriminability (b).

rising pair to form a broader similarity grouping of low-rising tunes. The low-falling/flat LLL tune stands alone—imitations of LLL are rarely misclassified. These results suggest a mapping of imitated tunes onto a similarity space with four clusters, described in terms of their holistic shape: high-rising, rise-fall, low-rising, and low-fall/flat. Within a cluster, tunes with distinct tone labels can be distinguished with lower accuracy, and are often misclassified for one another.

Turning to the perception results for the model tunes, Figure 3b plots estimates from the Bayesian model of the proportion of "same" responses given for a particular tune pair (corresponding to confusions in Fig. 2a, right), sorted from lowest to highest. In addition, the phonetic distance between the tunes in a given pair, calculated as the root-mean squared error between f0 trajectories of the model tunes (RMSE), is color-coded. RMSE is clearly related to perceptual discriminability. The model estimates shows that four tune pairs are discriminated at or below chance, based on 95% Credible Intervals that include or exceed 50% (chance), i.e., more "same" than "different" responses. The poorly discriminated tune pairs include the same tunes that are grouped together in the hierarchical clustering of the imitated production data: {HHH, HHL}, {HLH, HLL}, {LHL, LLH} and {LHH, LLH}.

We now have converging evidence for weak distinctions for four tune pairs, both in the perception of the model tunes, and in imitative productions of those tunes. Notably, all of these confusable tune pairs vary primarily in the f0 value at the end of the tune, seen in the model tunes (Fig. 1), and in the by-speaker average f0 trajectories of those tunes (not shown). Put differently, what the tunes in each pair have in common is their shape at the beginning of the nuclear word, including the f0 movement associated with the pitch accent.

Classifier confusions for some of the imitated tune pairs are lower than perceptual confusions of the model tunes for the same pairs: {LLH, LLL} and {LHH, LHL} (see Fig. 2b). This finding suggests that in imitating these tunes, speakers are enhancing relatively small F0 distinctions present in the model tunes. Conversely, classifier confusions for other imitated tune pairs are higher than perceptual confusions of the model tunes: {HHH, HHL} and {HLH, HLL}, which suggests that speakers are diminishing F0-based distinctions in imitating these tunes.

Which tune distinctions are robust? The hierarchical clustering of imitated tunes and the perceptual discrimination of model tunes alike indicate robust distinctions between most tunes. The most robust distinctions are between the tune groups that emerge from the hierarchical clustering analysis, i.e., high-rising tunes that start high and end higher {HHH, HHL}, rise-fall tunes {HLL, HLH}, low-rising tunes that end in a mid-level f0 {LHH, LLH, LHL}, and low-falling/flat tunes that start and end low {LLL}. Two of these clusters can be described in terms of their tonal specification: high-rising {HHX} and rise-fall {HLX}. The grouping of the low-rising cluster {LHL, LLH} does not neatly align with tonal specification, nor does the larger grouping of low-rising tunes that includes LHH.

4. Conclusions

This study tested distinctions in the perception and imitative production of 8 hypothesized nuclear tunes of American English. Tune pairs whose f0 trajectories are phonetically well separated are generally perceived and reproduced as distinct, while tune pairs that are closer in f0 space are more likely to be confused. Thus, we found similar results for classifier confusions and perceptual confusions for some tune pairs, but also divergent results, indicating that imitations sometimes enhance and sometimes minimize F0 distinctions in the model tunes. Classification accuracy of the imitative productions is highest for tune pairs that differ in holistic shape (high-rising vs. rise-fall), and poorest for tunes with the same shape that differ in (higher vs. lower) final f0. One limitation of this study is that we focus solely on f0; certain pitch distinctions may be perceived differently if co-varying with other cues such as duration and intensity. Broadening the cues under consideration in both human perceptual discrimination and machine classification will accordingly be a useful further direction.

Our results partially converge with findings from [11], suggesting that the differential scaling of tunes in a larger f0 space does not fully account for lost tune distinctions in that study. Major shape distinctions (e.g., low-rising vs. rise-fall) are maintained in both studies, while smaller differences in final f0 for tunes of the same shape are not. This finding raises questions about the categorical status of distinctions in holistic tune shape vs. final f0, and in the potential for each to convey distinctions in discourse meaning. We leave this for future research.

5. Acknowledgements

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

6. References

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

[2] Ladd, D.R. (2008). *Intonational Phonology.* Cambridge: Cambridge University Press.

[3] 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.

[4] Ladd, D. R. (1983). Phonological features of intonational peaks. *Language* 59: 721-759.

[5] 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*(2012), 1–49. https://doi.org/10.1515/lp-2012-0017

[6] 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. Retrieved from www.journalofspeechsciences.org

[7] Ladd, D. R., & Schepman, A. (2003). "Sagging transitions" between high pitch accents in English: Experimental evidence. *Journal of Phonetics*, 31,81–112.

[8] Shue, Y., Shattuck-hufnagel, S., Iseli, M., Jun, S., Veilleux, N., & Alwan, A. (2009). On the acoustic correlates of high and low nuclear pitch accents in American English. *Speech Communication.* https://doi.org/10.1016/j.specom.2009.08.005

[9] 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. https://doi.org/10.1080/01690961003599954

[10] Dainora, A. (2006). Modeling intonation in English: A probabilistic approach to phonological competence. In L. Goldstein, D. Whalen, & C. Best (Eds.), *Laboratory Phonology 8* (pp. 107-132). Berlin and New York: Mouton de Gruyter.

[11] Chodroff, E., & Cole, J. (2019). Testing the distinctiveness of intonational tunes: Evidence from imitative productions in American English. In *Proc. Interspeech 2019*, pp. 1966-1970.

[12] Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. *Proc. Interspeech 2012*.

[13] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter and H. Ney, "A comprehensive study of deep bidirectional LSTM RNNS for acoustic modeling in speech recognition," *Proc. ICASSP 2017,* pp. 2462-2466. doi: 10.1109/ICASSP.2017.7952599.

[14] Boersma, Paul & Weenink, David (2021). Praat v 6.1.48 retrieved June 4 2021 from http://www.praat.org/

[15] 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*.

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

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

[18] Sundberg, J. (1973). *Data on maximum speed of pitch changes*. Speech Transmission Laboratory Quarterly Progress and Status Report, KTH, Stockholm, Sweden. Retrieved from https://www.speech.kth.se/prod/publications/files/qpsr/1973/1973_14_4_039-047.pdf

[19] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[20] Bürkner P (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, **80**(1), 1–28. doi: 10.18637/jss.v080.i01.