A Comprehensive Framework for F_0 Estimation and Sampling in Modeling Prosodic Variation in Infant-Directed Speech

Meisam K. Arjmandi[1], *Laura C. Dilley*[1], *Matt Lehet*[1]

[1]Michigan State University, USA

khalilar@msu.edu, ldilley@msu.edu, lehetmat@msu.edu

Abstract

Accurate estimates of F_0 are essential for modeling how pitch variation is used as an informative cue to linguistic structure. Multiple challenges exist for estimation and valid statistical modeling of F_0 variation. First, certain speech styles, such as infant-directed speech, can involve dramatic pitch variation across utterances. Second, non-modal phonation can cause spurious F_0 values. Third, F_0 samples are not independent of one another, leading to issues with validity in applying generalized linear mixed effect models (GLMMs). To address these problems, we propose a comprehensive framework for accurate F_0 estimation and sampling to model prosodic variation. Our method involves segmentation of speech into utterances, followed by determination of speaker- and utterance-specific pitch range parameters. Regions of non-modal phonation are identified, ensuring that portions of speech leading to spurious F_0 values are rejected early. Next, F_0 stylization at the utterance level ensures robustness to microprosodic variation. Finally, F_0 turning points (e.g., local F_0 minima and maxima) are extracted; these are linguistically significant "control points" in F_0 contours connected by monotonic interpolations. This overall approach not only ensures accurate F_0 estimates, but critically overcomes the problem of non-independence of successive samples for valid statistical treatments within GLMMs.

Index Terms: Fundamental frequency estimation, F_0 contour stylization, generalized linear mixed effect models (GLMMs)

1. Introduction

Speech carries segmental and suprasegmental cues which provide listeners with a wide range of contextual and indexical information, such as the intended message, the speaker's age, gender, speech style, dialect, accent, emotional expression, and attitudes. Fundamental frequency (F_0) facilitates the perception of much of this information [1]. Listeners incorporate local and global F_0 fluctuations into perception of words, phrases, and utterances, using this variation to understand syntactic structure and meaning [2]–[4]. Even at an early age, dynamic use of F_0 in infant-directed speech (*i.e.*, motherese) is important to encourage infants' attention to speech, helping caregivers engage them and boosting their language development [5]. Refining the procedures in place to study this important acoustic cue is therefore a top priority in the field. In the present work, we propose a procedure for identifying reliable estimates of F_0 in utterances despite non-modal phonation, using these values in a stylization process that abstracts F_0 contours and allows the identification of linguistic control points. These control points (*e.g.*, local F_0 minima and maxima) provide independent observations of F_0 that overcome the non-independence problem that is a known obstacle for using F_0 data in generalized linear mixed effects models (GLMMs).

1.1. Challenges to F_0 Estimation

Despite its importance, accurate and reliable modeling of F_0 is still challenged by issues such as non-modal phonation, where speakers either intentionally or unintentionally introduce irregularities in voiced segments. These irregularities may result from disordered phonation [6]–[8], harsh or hoarse voicing [9], [10], or simply from intentional variations such as creakiness (glottalization) or breathiness [11]–[13]. These sources of variability in F_0 can negatively impact the accuracy of F_0 estimation. For example, the F_0 values extracted from these contexts often result in unwanted artifacts such as F_0 halving or F_0 doubling. Fundamental frequency estimation can also fail due to the presence of noisy components during breathy phonation [14]. Non-modal phonation during vocalization – whether from intentional or disordered sources – poses a major challenge to algorithms of F_0 extraction. Therefore, reliable estimation of F_0 must ideally include a strategy for pre-identification and exclusion of non-modal phonation from further analysis, reducing the burden on post-processing.

The accuracy of F_0 extraction also relies on a set of parameters that must be adaptively adjusted based on the acoustic qualities of the analyzed vocalization [15]. A robust autocorrelation method of F_0 extraction is implemented in *Praat* [16]. Accurate F_0 extraction in this software depends on aligning pitch parameters to the specific features of the analyzed speech. Vocalization patterns are prone to a wide range of variability due to anatomical and personal habits. Given this variability, identifying the range and variation in the minimum and maximum pitch must be specific to each speaker's voicing behavior.

The importance of pre-processing the speech signal to identify optimal speaker-specific values for these parameters during F_0 extraction was demonstrated by Keelan et al. [15]. They compared five F_0 extraction algorithms – auto correlation, cross correlation, sub-harmonic summation [16], the robust algorithm for F_0 tracking (RAPT) [17], and SWIPE' [18]. They showed the key role of speaker-specific F_0 parameter optimization in improving accurate F_0 estimation. Moreover, the F_0 dynamics across words and utterances during conversational speech varies within individual and across contexts. Therefore, it follows that an utterance-by-utterance and speaker-by-speaker parameter optimization strategy is required to appropriately deal with variability in pitch range, which is especially challenging in infant-directed speech. One way to overcome this issue is to identify and pre-define the optimum pitch parameters based on speaker's phonatory patterns and acoustical properties of speech contexts. This is a critical step in enhancing the reliability and validity of pitch

extraction. The role of this parameter optimization becomes more important when F_0 varies dramatically across stretches of time, such as the case of infant-directed speech.

1.2. Non-independence of successive F_0 values.

Whereas the technical challenges to F_0 extraction can be overcome procedurally, the extraction of raw F_0 values does not necessarily provide linguistically meaningful information. Crucially, these raw F_0 values are not independent observations [19]–[21]. Speakers adjust their vocal fold gestural patterns over multiple timescales, from short syllabic units to longer utterance-level intervals. This dynamic change in phonation creates various intonational structures, forming syllabic, lexical and phrasal constituents [2], [22]. The temporal variations of F_0 are usually characterized by two parameters (local minima or maxima timing and the change between consecutive high and low F_0 turning points), which provide substantial information for decoding a variety of linguistic, expressive, and organic information [2], [23].

The first parameter – the timing of local F_0 minima and maxima (i.e. turning points) where the slope of F_0 roughly changes from increasing to decreasing or vice versa – can be thought of as the identification of linguistically significant "control points". Considerable empirical evidence now suggests that the type, alignment, and scaling of F_0 turning points are broadly linked to distinctions in linguistic meaning [24]–[26]. These points are sparsely distributed in utterances and are connected by monotonic interpolation functions [27], [28]. Data reduction steps which prioritize identification of turning points and elimination of interpolations between them therefore is a valid means of addressing the problem of non-independence of successive F_0 samples, since these points reflect linguistic planning on the part of the speaker. The F_0 values that remain after such data reduction steps can therefore provide a window into dynamic communicative use of F_0; importantly, these values can be used in GLMMs due to the resolution of non-independence in successive F_0 samples, a significant contribution of the present work. The second parameter is the frequency differences between consecutive high and low F_0 salient turning points.

A reliable method for modeling F_0 fluctuation must reduce the continuous information in the F_0 contour (*i.e.*, F_0 stylization) to more linguistically relevant syllable level and utterance level information that models word and phrasal prominences. An effective F_0 stylization approach is an important step in identification of true minima and maxima as acoustic markers of vocal targets in the F_0 contour [29]. Such stylization is especially important in cases such as motherese, where F_0 dynamics directly impact infants' language development. By stylizing F_0 based on the local extremes within each syllable, and then identifying global extremes across utterances, we can reduce the continuous F_0 contour to a sequence of independent observations.

1.3. Using F_0 in Generalized Linear Mixed Effects Modeling

In order to accurately model the influence of F_0 variation on various psycholinguistic variables, it is also important to use appropriate statistical analyses. Generalized linear mixed effects models (GLMMs) have recently been promoted as providing statistical models that can account for extraneous variance from multiple random sources, as well as fixed effects of independent variables [30], [31]. Barr thoroughly discussed how treating non-independent observations as independent can result in underestimation of standard errors, which subsequently can bring inflated Type I error [31]. To conduct a valid statistical analysis without encountering this issue, a GLMM can be built on F_0 minima and maxima as independent observations instead of using the non-independent measurement of successive F_0 samples on the millisecond scale [31]. GLMMs can be used to model F_0 variations both at the syllabic timescale, based on extrema derived from stylization of raw F_0 within each vowel phones, and identifying *turning points* on the stylized F_0 track within each utterance as independent observations. These minimum and maximum F_0 values (extrema) that correspond to turning points in the F_0 contour are linguistically important.

This study proposes an extensible framework that provides accurate and reliable F_0 extraction and a F_0 contour modeling method by addressing the issues identified above. The present work proposes applying GLMM on the independent observations derived from local (syllable-level) and global (utterance-level) turning points, providing a more precise model of F_0 over time. The F_0 extraction in the present study is based on the idea that both local and global patterns of F_0 alternations carry substantial linguistic information essential for decoding the utterance meaning. Our proposed framework represents both macroprosodic and microprosodic structures in utterances which are important in carrying linguistic information [39]. This approach is critical to a comprehensive understanding of how intonational patterns in motherese speech influence infants' language outcome.

2. Proposed Framework for F_0 Characterization

To robustly and reliably characterize temporal fluctuation of F_0, we developed a new framework that addresses the challenges discussed above. The block diagram of our proposed framework is illustrated in Figure 1. This method was developed as part of an ongoing study with the goal of investigating the relationship between prosodic characteristics of speech directed to infants with hearing loss and their language outcomes. This is a semi-automatic method that uses pre-defined F_0 parameters and prelabeled speech utterances to model intonational pattern at the level of syllable and utterance.

2.1. Utterance Segmentation and F_0 Parameter Optimization

Our proposed F_0 extraction method starts by excluding – by hand – regions of speech likely to generate spurious F_0 values. The speech waveform and spectrogram are viewed along with a superimposed visual display of the F_0 curve. Regions of modal phonation are marked in *Praat TextGrid* with 'x' for inclusion. Regions of non-modal phonation are identified through visual inspection of the waveform, including creak and diplophonia, and by virtue of no 'x' label are excluded. Further, four parameters – *pitch floor, pitch ceiling, silence threshold, and voicing threshold* – are adjusted within each vocalized region to provide a speaker-specific and acoustic-context-specific optimization for each F_0 parameter on an utterance-by-utterance basis. Analysts visually inspect F_0 curves for each utterance in comparison with auditory impressions, iteratively adjusting the four named parameters in

order to ensure a smooth and accurate F_0 curve. Any changes from *Praat*'s default settings that are required to ensure a smooth and accurate curve, *e.g.*, adjustment to the pitch ceiling of 600 Hz, are explicitly marked in the tier. The result is a set of labels in a *Praat* tier which are used by a script to exclude portions of speech expected to inaccurate F_0 estimates (e.g., non-modal phonation), and/or to adjust *Praat*'s default autocorrelation parameters, resulting in a set of highly accurate F_0 values.

2.2. Outlier Removal

By eliminating portions of speech with non-modal phonation in pre-processing steps as outlined above, instances of spurious F_0 values are dramatically reduced, enhancing accuracy overall. Additional robustness can be achieved by further post-processing to remove outlier F_0 values. It was shown that speaker-specific raw F_0 values are not necessarily normally distributed and are often positively skewed [34]. This implies that incorporation of an outlier removal method that is tailored to the speaker-specific distribution of F_0 values is useful to identify and exclude spurious values. This post-processing distribution-specific outlier removal approach is implemented in *Matlab 9.3.0.* (The Mathworks, http://www.mathworks.com). Spurious values were identified as values more than three scaled *median absolute deviation* (MAD) units from the distribution median as a robust measure of dispersion [35]. The scaled MAD is defined as $SMAD_i = c_i * median(|x_i - median(x_i)|)$ where x is the sequence of F_0 values for each speaker i. The scaled parameter is calculated as in [35]. The number of spurious raw F_0 values detected in this step was small since the previously applied F_0 parameter optimization method strongly reduced the chances of inaccurate F_0 values.

2.3. F_0 Stylization

F_0 stylization has been used to simplify raw F_0 fluctuations over time and instead represent linguistically salient tonal information. The framework developed in the current study proposes two sequential steps. The first stylizes the F_0 contour within each syllable based on the minima and maxima within the syllable. The second step recombines these stylized contours and identifies minima and maxima across the utterance. This approach reduces continuous F_0 contours into independent observations that reflect non-monotonic minima and maxima within and across utterances.

Our proposed F_0 contour stylization at the syllable level first uses manually identified syllables within utterances [36] with breaks between utterances taken as pauses of 250 *ms* or greater. The F_0 contour within each syllable is stylized using an adaptive scheme. If the variation of F_0 within each syllable is monotonic, the contour is stylized using a linear fit [36], otherwise a 2^{nd}-order spline function is used to stylize F_0 within each syllable as proposed by *Momel* [37]. Then, the minimum and maximum, derived from the stylized F_0 contour within each syllable, are calculated and used as estimates of syllable-level F_0 variations to fit a 2^{nd}-order spline function. The turning points on the globally fitted syllable-based F_0 contour are used as independent observations of F_0 variation for each utterance. This stylization approach is designed to appropriately model highly fluctuating F_0 contours in infant-directed speech [37], taking into account the structure within syllables, while identifying trends across syllables.

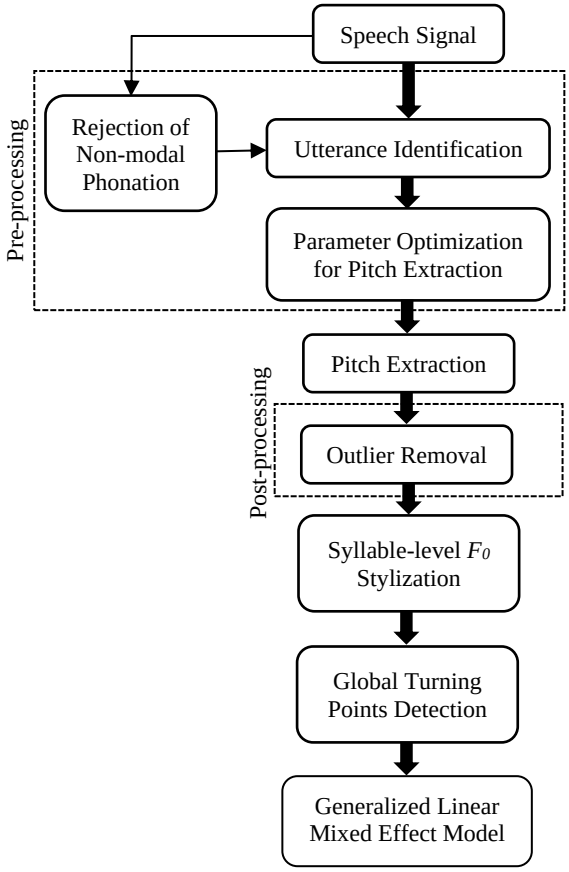

Figure 1: *The proposed framework for characterization of F_0 variation over time and modeling the fixed effects of F_0 variation on infants' language learning outcome.*

2.4. Application in a Generalized Linear Mixed Effects Model (GLMM)

The utterance-level turning points resulting from the aggregated stylized F_0 contours within syllables can serve as independent observations to build GLMMs. These utterance level extrema obtained from the proposed algorithm reflect syllabic level detail but can be used to model fixed effects of F_0 dynamics in infant-directed speech or adult-directed speech with subject-level random effects for each talker. This is an ongoing project where we will use the comprehensive framework presented here to investigate how local and global variation in timing, distribution, and magnitude of F_0 within and across utterances correlates with language development in children with hearing loss.

3. Discussion

Fundamental frequency (F_0) is an important acoustic feature that cues linguistic and indexical information in speech. Tracking the fluctuation of F_0 in continuous speech to model prosodic structure faces challenges that may negatively affect the reliability and validity of the estimated parameters from F_0 contours [13], [15], [34], [38]. The present work proposes a comprehensive framework for modeling local and global patterns of F_0 dynamics by providing methodological solutions

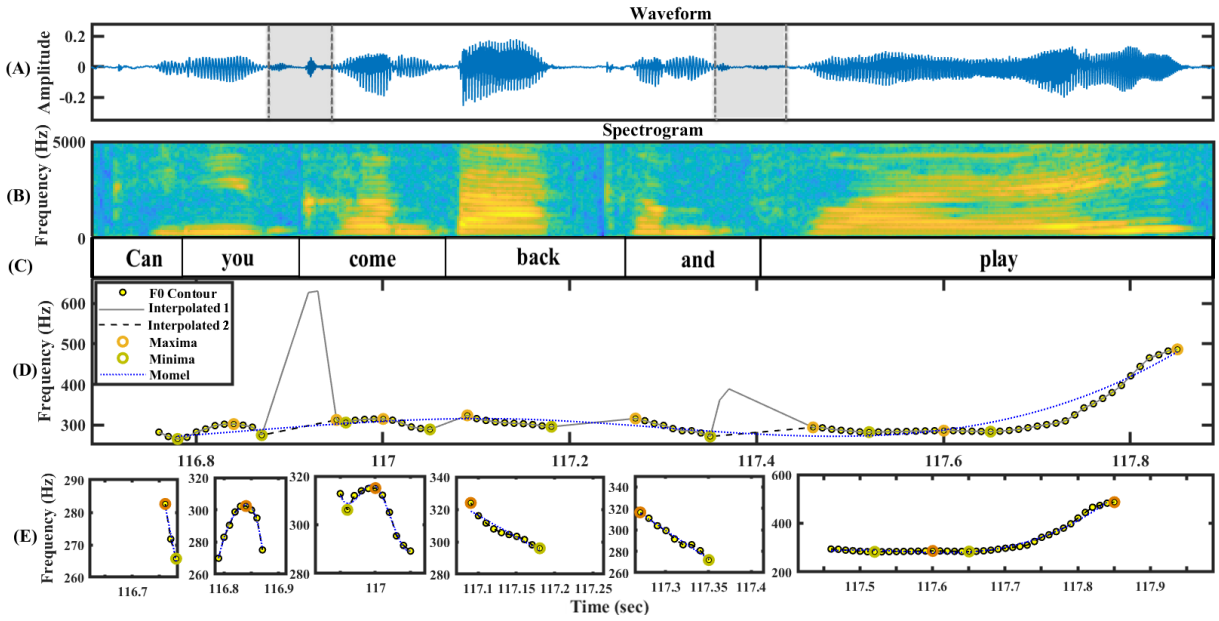

Figure 2: *A schematic of macroprosodic (i.e., post-lexical) and microprosodic (i.e., syllabic) F_0 variation in a typical sentence spoken by a mother to her child using infant-directed speech. (A) Waveform of the sentence. (B) Spectrogram of the sentence. (C) The manual transcript of words and their boundaries. (D) The macroscopic variation of original F_0 values transformed in Hz (black circles with yellow facecolor), the interpolated F_0 contour (dashed and gray line), local minima (green circles) and maxima (orange circles) of F_0 contour, and Momel F_0 contour stylization (blue dotted line). (E) Microprosodic variation of F_0 within each word and the detected local minima and maxima.*

to address these challenges. In this framework, the likelihood of spurious F_0 outcomes is reduced by identification and exclusion of creaky, breathy, and other non-modal phonations. The issue of parameter optimization is addressed by proposing a speaker-specific and context-specific F_0 parameter adjustment scheme. An outlier removal procedure tailored to speaker-specific F_0 distribution assures thorough artifact removal in the F_0 extraction. These pre-processing steps ensure reliable extraction of F_0 from complex acoustic environments despite previously identified challenges in the literature.

Our proposed method for modeling F_0-related prosodic structure in sentences is based on evaluating F_0 fluctuation at syllabic and utterance level specificity within the same analysis [3]. In this framework, F_0 contour stylization is performed on short syllable length time scales, and then characterized at longer utterance level time scales, capturing both local and global variations in F_0. By identifying non-monotonic minima and maxima points at these time scales we can characterize the magnitude and timing of F_0 dynamics in a discrete set of independent observations. These independent observations of F_0 fluctuation allow us to model the relationship between F_0 dynamics and language development using GLMMs. We are specifically interested in modeling how F_0 differences between infant-directed and adult-directed speech impact language-learning outcomes in infants with cochlear implants.

Despite solving many challenges in F_0 analysis, the current procedure relies on hand-coded syllable boundaries. An automatic syllabification procedure would streamline analysis, but highly co-articulated word boundaries in natural speech, as shown by "*can you …*" segments [39], [40] (panel (B) and (C) of Figure 2) demonstrate one challenge in automatic vowel segmentation. Elongated speech stretches such as "… *play*" in Figure 2 also makes this automatic detection of vowel phones and syllable centers harder.

4. Conclusion

Despite advances in methods of F_0 characterization, there are still some critical challenges that need to be addressed. These challenges are particularly apparent when analyzing speech with large dynamic range of F_0 variation such as in motherese. The methods used in the field are not designed to deal with such speech, but accurate modeling of F_0 as it varies over time is critical to understanding the role of prosodic cues delivered by motherese speech in infant's language development. To this end, it is strategically important to understand whether infants benefit from prosodic cues conveyed by global rising and falling in F_0 contour. By identifying independent observations of F_0 at the level of syllables and utterances, we can begin to understand how F_0 structures language learning.

Our proposed framework provides a comprehensive F_0 contour parametrization approach that allows F_0 data to be analyzed using GLMMs. This study provides a framework through which the effect of tonal variation in infant-directed speech on language outcome can be reliably modeled. The outcome is also translational to a clinical application of speech and language development.

5. Acknowledgements

Research reported in this paper is supported by National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health (NIH) under award number of R01DC008581-06A1.

6. References

[1] C. J. Chen, H. Li, L. Shen, and G. Fu, "Recognize tone languages using pitch information on the main vowel of each syllable," *Int. Conf. Acoust. Speech Signal Process.*, no. 1, pp. 61–64, 2001.

[2] R. A. Knight, "The shape of nuclear falls and their effect on the perception of pitch and prominence: Peaks vs. plateaux," *Lang. Speech*, vol. 51, no. 3, pp. 223–244, 2008.

[3] S.-A. Jun, "Prosodic typology: by prominence type, word prosody, and macro-rhythm," in *Prosodic Typology II: The Phonology of Intonation and Phrasing*, Oxford: Oxford University Press, 2014, pp. 520–540.

[4] L. Frazier, K. Carlson, and C. Clifton, "Prosodic phrasing is central to language comprehension," *Trends Cogn. Sci.*, vol. 10, no. 6, pp. 244–249, 2006.

[5] H. Fernald and P. Kuhl, "Acoustic determinants of infant preference for motherse speech," *Infant Behav. Dev.*, vol. 10, pp. 279–293, 1987.

[6] V. Parsa and D. G. Jamieson, "Identification of Pathological Voices Using Glottal Noise Measures," pp. 469–486, 2000.

[7] M. K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, and A. Moqarehzadeh, "Identification of Voice Disorders Using Long-Time Features and Support Vector Machine With Different Feature Reduction Methods," *J. Voice*, vol. 25, no. 6, pp. e275–e289, Nov. 2011.

[8] M. K. Arjmandi and M. Pooyan, "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine," *Biomed. Signal Process. Control*, vol. 7, no. 1, pp. 3–19, Jan. 2012.

[9] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic Correlates of Breathy Vocal Quality," *J. Speech Lang. Hear. Res.*, vol. 37, no. 4, p. 769, 1994.

[10] T. Yumoto, E., Gould, W. J., & Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Am.*, vol. 71, no. 6, p. 1544–1550., 1982.

[11] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels as a function of prosodic structure," *J. Phon.*, vol. 24, no. 4, pp. 423–444, 1996.

[12] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *J. Phon.*, vol. 29, no. 4, pp. 407–429, 2001.

[13] O. Babacan, T. Drugman, N. D 'alessandro, N. Henrich, and T. A. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds comparative study of pitch extraction algorithms on a large variety of singing sounds," pp. 1–5, 2013.

[14] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. August, pp. 1973–1976, 2011.

[15] K. Evanini and C. Lai, "The importance of optimal parameter setting for pitch extraction.," *J. Acoust. Soc. Am.*, vol. 128, no. 4, p. 2291, 2010.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 6.0.37," 2018.

[17] D. Talkin, W. B. Kleijn, and K. K. Paliwal, "A Robust Algorithm for Pitch Tracking (RAPT)," *Speech Coding and Synthesis, The Eds. Amsterdam, Netherlands:Elsevier*. pp. 495–518, 1995.

[18] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[19] K. Silverman *et al.*, "TOBI: A Standard for Labeling English Prosody," *Second Int. Conf. Spok. Lang. Process.*, no. October, pp. 867–870, 1992.

[20] C. d'Alessandro and P. Mertens, "Automatic pitch contour stylization using a model of tonal perception," *Comput. Speech Lang.*, vol. 9, no. 3, pp. 257–288, 1995.

[21] A. Origlia, G. Abete, and F. Cutugno, "A dynamic tonal perception model for optimal pitch stylization," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 190–208, 2013.

[22] H. Fujisaki and T. Kawashima, "The Roles of Pitch and Higher Formants in the Perception of Vowels," *IEEE Trans. Audio Electroacoust.*, vol. 16, no. 1, pp. 73–77, 1968.

[23] X. S. Shen, M. Lin, and J. Yan, "F0 turning point as an F0 cue to tonal contrast: A case study of Mandarin tones 2 and 3," *J. Acoust. Soc. Am.*, vol. 93, no. 4, p. 2241, 1993.

[24] P. Prieto, M. D'Imperio, and B. Gili Fivela, "Pitch accent alignment in romance: Primary and secondary associations with metrical structure," *Lang. Speech*, vol. 48, no. 4, pp. 359–396, 2005.

[25] M. Grice, D. R. Ladd, and A. Arvaniti, "On the place of phrase accents in intonational phonology," *Phonology*, vol. 17, no. 2, pp. 143–185, 2000.

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

[27] L. C. Dilley and C. C. Heffner, "The role of f0 alignment in distinguishing intonation categories: evidence from American English," *J. Speech Sci.*, vol. 3, no. 1, pp. 3–67, 2013.

[28] D. R. Ladd and A. Schepman, *"Sagging transitions" between high pitch accents in English: Experimental evidence*, vol. 31, no. 1. 2003.

[29] P. Mertens, "The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model," *Proc. 2nd Int. Conf. Speech Prosody*, pp. 549--552, 2004.

[30] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *J. Mem. Lang.*, vol. 68, no. 3, pp. 255–278, 2013.

[31] D. J. Barr, "Analyzing 'visual world' eyetracking data using multilevel logistic regression," *J. Mem. Lang.*, vol. 59, no. 4, pp. 457–474, 2008.

[32] S. Chen, C. Zhang, A. G. McCollum, and R. Wayland, "Statistical modelling of phonetic and phonologised perturbation effects in tonal and non-tonal languages," *Speech Commun.*, vol. 88, pp. 17–38, 2017.

[33] E. D. Thiessen, E. A. Hill, and J. R. Saffran, "Infant directed speech facilitates word segmentation," *Infancy*, vol. 7, no. 1, pp. 53–71, 2005.

[34] M. Lennes, M. Stevanovic, D. Aalto, and P. Palo, "Comparing pitch distributions using praat and r," *Phonetician*, vol. 111, pp. 35–53, 2016.

[35] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *J. Exp. Soc. Psychol.*, vol. 49, no. 4, pp. 764–766, 2013.

[36] S. Ravuri and D. P. W. Ellis, "Stylization of pitch with syllable-based linear segments," in *Acoustics, Speech and Signal Processing (ICASSP 2008)*, 2008, pp. 3985–3988.

[37] D. Hirst and R. Espesser, "Automatic Modelling Of Fundamental Frequency Using A Quadratic Spline Function," *Travaua de l'Institut de Phonétique d'Ait*, vol. 15, pp. 71–85, 1993.

[38] H. Fujisaki and T. Kawashima, "The Roles of Pitch and Higher Formants in the Perception of Vowels," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, no. 1, pp. 73–77, 1968.

[39] L. Dilley, M. K. Arjmandi, Z. Ireland, C. Heffner, and M. Pitt, "Glottalization, reduction, and acoustic variability in function words in American English," *J. Acoust. Soc. Am.*, vol. 140, no. 4, pp. 3114–3114, 2016.

[40] L. Dilley, M. K. Arjmandi, and Z. Ireland, "Spectro-temporal cues for perceptual recovery of reduced syllables from continuous, casual speech," *J. Acoust. Soc. Am.*, vol. 141, no. 5, pp. 3700–3700, May 2017.