Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Prosodic variability in marking remote past in African American English

Kristine M. Yu, Alessa Farinella

University of Massachusetts, Amherst

krisyu@linguist.umass.edu, afarinella@umass.edu

Abstract

This paper explores variability in the fundamental frequency (f0) of utterances containing the remote past marker BIN in African American English, which has been described as having higher f0, intensity and duration relative to preceding material, and reduced f0 following, though with some interspeaker variability (Green et al. 2022). Here we re-analyze data from Green et al. (2022) to characterize the space of possible phonetic realizations of BIN utterances. We computed the 90th percentile f0 value in pre-/on-/post-BIN regions to create a 3-point "topline" f0 shape profile of the utterance (Cooper & Sorensen 1981) and performed time series clustering and principal components analysis (PCA). Two clusters were identified, one with higher f0 on BIN and lower f0 post-BIN, and one with lower f0 on BIN and higher f0 post-BIN. Results from PCA indicate speakers vary along two dimensions: one relating to pre-BIN f0 and one to post-BIN f0. Both dimensions were tied to f0 height on BIN, demonstrating the role that global aspects of the contour play in the variability. We show how the topline representation of f0 contour shape is robust to missing values and uncontrolled sentences and thus useful for naturalistic speech.
Index Terms: intonation, African American English, variability

1. Introduction

In African American English (AAE)[1] there are multiple types of "been" used for marking tense and aspect, including remote past *BIN* (orthography used by linguists) and auxiliary perfect *been* (see [1]). While they are string identical, they differ in meaning and prosody. Past work on prosody has compared auxiliary been used in perfect contexts (also present in Mainstream American English, or MAE), and remote past BIN used to situate an event in the remote past or for habitual actions [2]. The remote past BIN is not present in MAE, but a similar meaning can be expressed with a verb(s) + adverb or adverbial phrase [1]. Two pronunciations of the remote past BIN are shown in Figure 1, both elicited with the same context. The remote past BIN has been referred to as 'Stressed BIN' [3, 4], as it is often realized with prominence on 'BIN.' Acoustically, it has been described as having high fundamental frequency (f0), intensity and duration on BIN and compressed f0 range after BIN [2, 5].

Less attention has been paid to variability in the realization of BIN utterances across speakers, which can be clearly seen by comparing the two pitch tracks in Figure 1. The top pitch track shows a clear f0 peak on BIN, while the bottom shows a higher f0 before BIN and a much smaller peak on BIN (also noted in [5] and [6]).

[1]Here we adopt the term African American English to refer to a variety of English that has set syntactic, phonological, semantic, pragmatic, and lexical patterns that are intertwined with structures of Mainstream American English (MAE) [1]. It is sometimes referred to as African American Language or African American Vernacular English, among other names.

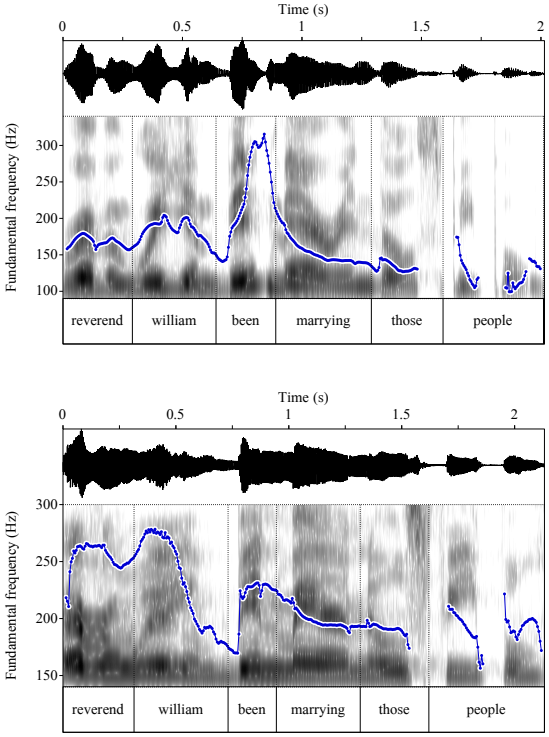

Figure 1: *Pronunciations of the remote past BIN from two different speakers, la09 (top) and la04 (bottom), elicited in the same context. The sentences could be paraphrased in MAE as: 'Reverend Williams has been marrying those people for a long time (and is continuing to do so)'*

The fact that this degree of variability exists despite a common remote past meaning raises interesting questions about the ways the same intonational category may be realized phonetically. Several past studies have looked at variation in AAE (e.g., [7]), but much work on prosodic variation in AAE has focused on comparisons with MAE (e.g., [8], [9], [10], [11]), rather than considering variability with respesct to the rest of the intonational phonology of AAE. In intonation research more generally, variation has been less explored than for segmental contrasts [12], [13], [14]. Looking at variation over the course of an utterance brings up specific challenges that are often not at issue when investigating segmental variation, especially when it comes to more naturalistic speech. These include missing f0 values, and differences in sentence length and segmental content.

This paper has two main goals. The first is to take a deeper look at prosodic variation in BIN utterances, a necessary first step toward understanding the relation between prosody and

10.21437/SpeechProsody.2024-215

the remote past aspectual marker in AAE. Characterizing this variation will make it possible to explore conditioning factors. The second goal is to demonstrate a viable methodology for exploring intonational variation that addresses the challenges inherent in working with naturalistic data, a point often overlooked, as intonation research often deals with carefully controlled data. We test an acoustic, data-driven method to capture variation in f0 contour shape by generating f0 "toplines" [15], which captures overall f0 contour shapes while being robust to missing f0 values (due to segmental perturbations or interruptions/disfluencies) and is generalizable across different sentences of different lengths. We demonstrate why other data-driven methods including time series clustering and functional principal components analysis, which have become increasingly popular in work on intonation ([14],[16], [17], [18], [19], [20] and many others), face challenges when applied to naturalistic data. We apply our topline method to a multispeaker dataset of BIN utterances, which allows us to then make use of clustering and principal components analysis. Using these methods, we discover several patterns in prosodic variation over the course of BIN utterances, which enriches our understanding of the prosodic realization of the remote past in AAE.

2. Materials and Methods

2.1. Materials

Data come from 8 adult members of an AAE-speaking community in southwest Louisiana (5 female, 3 male), previously analyzed in [5], which also describes the stimuli, procedures, and speaker demographics in full detail. In brief, remote past BIN and perfect been utterances were elicited with written prompts, with situational context presented auditorily and visually. Recordings were segmented and forced aligned using the Montreal Forced Aligner [21]. A linguistically-trained native speaker from the community classified each elicited utterance in isolation as a BIN or been perfect and also judged each utterance for acceptability in the situational context provided.

Only the 311 tokens unambiguously classified as BIN and judged acceptable were included for analysis here. While this data set is small, it is still currently the largest set of recorded BIN utterances available. The Corpus of Regional African American Language (CORAAL) [22]—now with over 160 hours of recordings—has become the go-to data source for much acoustic work on AAE in the past five years. But [5] found only 20 instances of BIN constructions in the entire corpus, potentially both because the semantic conditions required for a BIN construction rarely occur in sociolinguistic interviews, [4, p. 99] and also because, even if the semantic conditions were met, speakers chose to use an adverbial form like "for a while" instead of BIN 87% of the time.

2.2. Methods

F0 was extracted using Praat's autocorrelation algorithm [23], using the same speaker-specific f0 floor and ceiling values as [5] and otherwise default settings. F0 was extracted at 10ms intervals for full f0 contours; the 90th percentile f0 value was extracted from each word for toplines. The 90th percentile value rather than the f0 maximum was chosen for robustness against wide f0 excursions from segmental perturbations—a common strategy in large-scale f0 processing and the automatic detection of f0 range [24]. All further data analysis was done in R [25], and plots were drawn with `ggplot2` [26].

For full f0 contours, missing f0 values were trimmed or replaced by estimates computed from cubic spline interpolation using the `imputeTS` R package [27] (also used for f0 by [18]), see Sec. 3.1. We computed the topline of a BIN f0 contour as the 3-point sequence of the maximum 90th percentile f0 value over all words in each of the pre-BIN, on-BIN, and post-BIN regions, see Fig. 2. The word immediately following BIN was also included in the on-BIN region because there was sometimes peak delay on BIN. Missing values could be and were ignored for the topline computations. F0 data from function words *a*, *the*, *at*, *for* and *to* were excluded, since they were frequently all missing values, or impacted by segmental perturbations, like in f0 at the beginning of the [p] in *people* in Fig. 2.

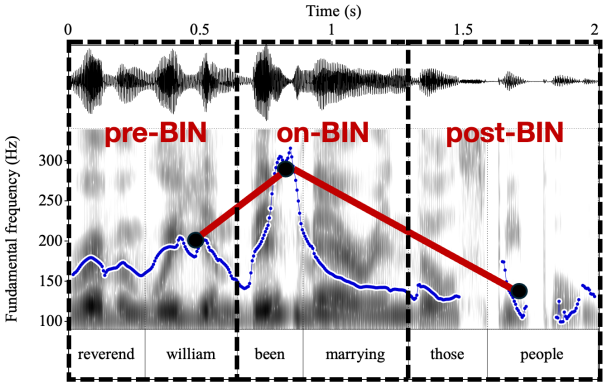

Figure 2: *Schematic of topline over highest peaks in pre-BIN, on-BIN, and post-BIN region for la09 example, Fig. 1.*

Binning the f0 contour into the pre-, on-, and post-BIN regions for the topline effectively time-normalized across utterances: every utterance's f0 contour was summarized as 3 points, regardless of how long the utterance or any of the BIN regions were. Similarly, BIN was "registered" as a landmark [28, §3.3] in the full f0 contour by aligning time courses at both the onset and offset of BIN, see the two vertical dividing lines in Fig. 8. Both full f0 contours and toplines were log-transformed and by-token mean centered.

Due to the limited amount of data available, we focused on visualization and exploratory data analysis rather than inferential statistics: (i) time series clustering and (ii) principal components analysis (PCA). These methods allowed us to explore, respectively: how can we group the space of BIN topline realizations? And what key properties (principal components) characterize the range of BIN topline variability? We performed partitional clustering based on Euclidean distance using `dtwclust` [29]. We tested having 2-4 clusters, and all cluster validity indices indicated 2 was the optimal number. PCA was computed using the `prcomp` function, and visualizations were aided by `factoextra` [30]. We also did clustering on the full f0 contours, for comparison, with the same settings. For full details, see the OSF repository at `https://osf.io/7qnvk/`.

3. Results

3.1. Distribution of missing values

Every single one of the 311 BIN f0 contours contained missing values (NAs)—altogether, 15,442 (23% of extracted f0 values). Thus, if we had followed the common strategy of omitting contours with missing values, we would have had no data left to analyze. Table 1 shows four key sources of missing values and how we handled them for the full f0 contour. We trimmed

Table 1: *Sources of missing values, percentage of missing values due to source, and distribution and strategies for handling them for full f0 contours*

Source	%	Strategy for handling
Utt-final plurals	39.9	Remove trailing NAs
the	6.7	Remove leading NAs, impute NAs
BIN	5.3	Impute NAs
Silence	3.9	Remove trailing/leading, impute
Other	44.2	Impute NAs

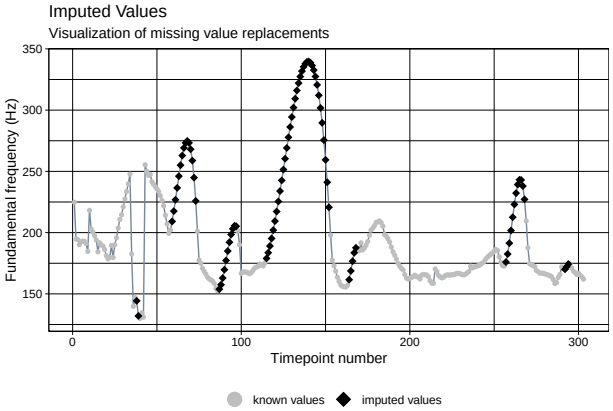

Figure 3: *Wild cubic spline imputation of f0 values over missing value region from utterance-medial [s] followed by 230-ms (fluent) pause, ≈ timepoints 100-150. Imputed values are black.*

utterance-initial ("leading") or -final ("trailing") strings of missing values from the f0 contour. This trimming handled at least some of the missing values from utterance-final words that happened to be s-final plurals (40% of NAs), e.g., *containers*, *winners*, and instances of *the* that were utterance-initial (17/47 BIN items). It also helped handle some silent intervals, e.g., modal stimuli included an utterance-initial "Aw" for naturalness, which speakers sometimes followed with a pause; we trimmed off "Aw", and that enabled us to trim off NAs in the pause as leading NAs.

Removing leading/trailing NAs eliminated 60% of the missing values (leaving 9,337) and reduced missing values from 288/311 tokens, but still left 308 tokens with missing values. The remaining, utterance-medial missing values had to be replaced with estimates (imputed with via imputeTS [27]) to maintain continuity of the f0 time series. Fig. 3 shows how wildly imputation can behave over a voiceless interval [s] followed by a pause (the giant spike in black around points 100-150) between *The maintenance workers* and *BIN*, as well as examples of imputation over other voiceless [s] regions. The imputed f0 spikes in the figure also highlight that handling missing values does not handle large f0 excursions due to segmental perturbations, which then greatly affect imputation downstream.

3.2. Topline results

Unlike the full f0 contours, the toplines were robust to missing values from voiceless regions, segmental perturbations, and silences and easily generalized across different sentences with different lengths, words, stress positions, etc. Median toplines are shown in Fig. 4 across speakers and for each individual

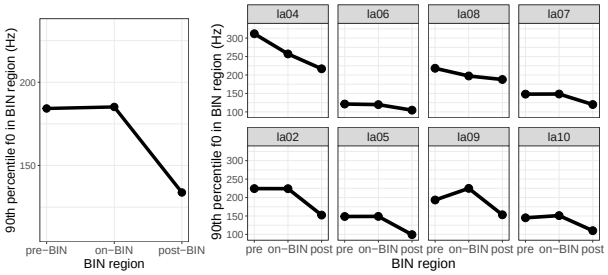

Figure 4: *Median topline across speakers (left) and for each speaker (right)*

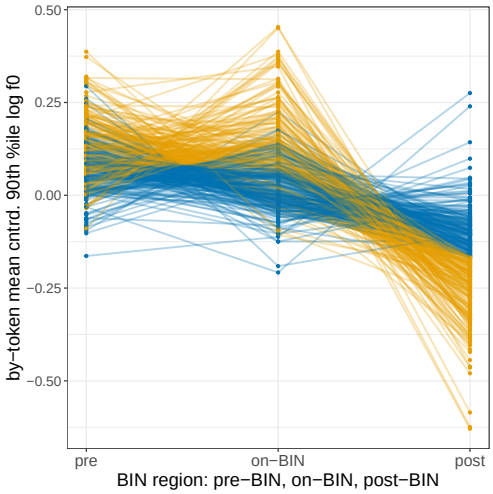

Figure 5: *Topline tokens, color coded by cluster. "little BIN": darker, in blue; "big BIN": lighter, in orange.*

speaker. Aggregating the data like in Fig. 4 obscures the presence of BIN contour realizations like Fig. 1, bottom.

The topline tokens split into two clusters that we nicknamed "big BIN" and "little BIN", shown in Fig. 5. The "big BIN" cluster (lighter, in orange) is characterized by higher on-BIN f0 and lower post-BIN f0. The "little BIN" cluster (darker, in blue) is characterized by a lower on-BIN f0 and a higher post-BIN f0. The distinction between the two clusters in terms of pre-BIN f0 is not as clear. Fig. 6 shows toplines by individual speaker, also color-coded by cluster like in Fig. 5. While the top row of speakers tended to produce "little BINs", the bottom row of speakers tended to produce "big BINs". However, all speakers produced tokens in both clusters.

PCA results yielded two principal components (PCs) that accounted for 99% of variance in toplines, see Fig. 7. PC1 accounted for 68% of the variance and correlated most strongly with lower post-BIN and to a lesser degree with higher on-BIN f0. PC2 (32% of variance) correlated most strongly with lower pre-BIN and to a lesser degree with higher on-BIN f0. In other words, the primary dimension of variation (PC1) used by the speakers for BIN toplines was to push down post-BIN f0 while simultaneously raising the on-BIN peak. Independently, the secondary dimension of variation was to push down pre-BIN f0 while simultaneously raising the on-BIN peak. The simultaneous lowering outside the BIN region and raising in the BIN region characteristic of both PCs is reminiscent of a seesaw.

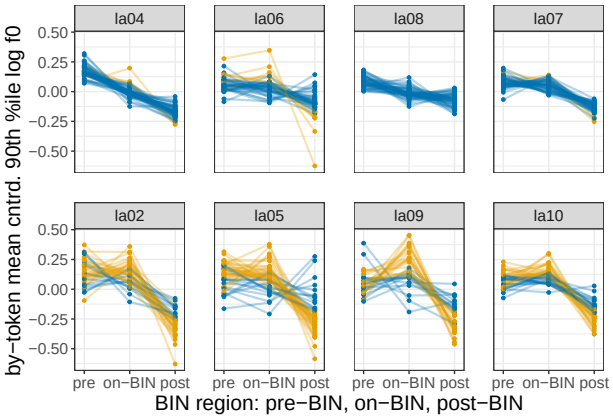

Figure 6: *Topline tokens by speaker, color coded by cluster assigned as in Fig. 5, "big BIN": lighter, in orange; "little BIN": darker, in blue.*

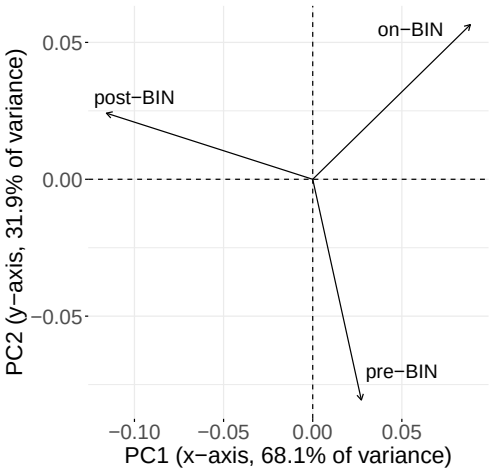

Figure 7: *Correlation circle for PCs. PC1 was highly negatively correlated with post-BIN f0, and PC2 with pre-BIN f0. Both PCs were moderately positively correlated with on-BIN f0.*

3.3. Full f0 contour results

Clustering over the full f0 contours results into 291 tokens falling into one cluster and only 20 in the other. Fig 8 shows the distribution of tokens from each cluster for each individual speaker. Unlike the topline characterizations in Fig. 6, it is not clear what underlies the division into clusters nor that there are speaker-specific tendencies for BIN realization.

4. Discussion and Conclusion

The empirical contribution of this paper is a a first acoustic exploration of prosodic variability in the realization of remote past BIN utterances in African American English, a variety of English that is still relatively prosodically underdescribed. Two distinct clusters emerged from time series clustering: one with higher f0 on BIN and lower f0 post-BIN, and another with higher post-BIN f0 and a smaller peak on BIN. Speakers tended to produce realizations from mostly just one of the clusters. What might underlie speaker choice of BIN rendition is yet unclear. PCA results also highlight the main dimensions manipulated by speakers in generating variability in the realization of

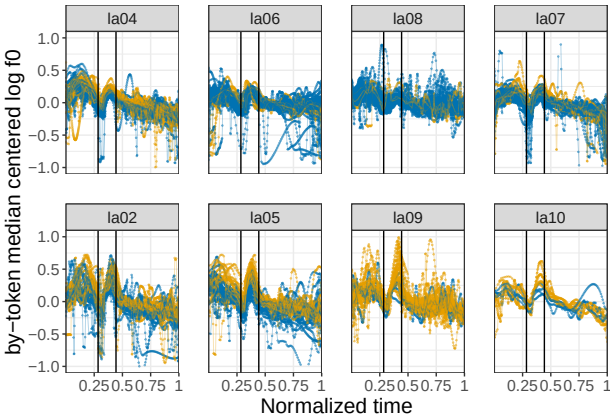

Figure 8: *Division of full f0 contours into two clusters by speaker.*

BIN constructions: a choice to fall to a lower f0 after BIN is tied to reaching a higher peak on BIN. Our findings highlight that aspects of the overall contour, and not just f0 scaling on BIN alone, is involved in variation in BIN utterances, although past descriptions of BIN have emphasized high f0/tone on just BIN itself. While [5] found, in aggregate, that BIN utterances have higher f0 peaks on BIN than in the pre- or post-BIN regions, a substantial number of individual tokens examined here do not necessarily conform to this pattern, i.e., some proportion of "little BIN" realizations. The acoustic methods alone cannot provide a decision on whether "little BIN" variants like la09's bottom utterance in Fig. 1 might be phonologically or otherwise categorically distinct from "big BIN" variants, or other "little BIN" realizations that don't start out with such a high f0. But the topline cluster labels and PC parameterizations can facilitate identifying BIN variants like Fig. 1 and uncovering what conditioning factors give rise to them.

Our methodological contribution is to draw attention to challenges with extending shape-based methods such as time series clustering, functional PCA, and GAMMs to more naturalistic data and characterizing f0 shape patterns extending over much longer windows than a syllable. The algorithms used in these methods typically assume/expect time courses without any missing values. But more naturalistic, uncontrolled data—as well as longer analysis windows—all increase the chances of having missing f0 values in the f0 contour due to voiceless intervals, silence, etc. We showed that while it is possible to remove or impute missing values, imputations may introduce improbable jumps in the f0 contour. Removing trailing/leading missing values can affect duration/time normalization. It is also important to consider the sources of missing values. When missing values arise from silence in a mid-utterance juncture, no interpolation of f0 values over the silence makes sense. But if shape components spanning across the juncture are of interest, then the missing values need to be filled in for the shape-based methods. We demonstrated that a viable alternative is hearkening back to 1970s-1980s low temporal resolution representations of downtrend shapes such as the "topline" over f0 peaks, which also has the benefit of easily generalizing across different sentences with different lengths, stress positions, etc., due to its sparsity of sampling.

5. References

[1] L. J. Green, *African American English: a linguistic introduction*. Cambridge: Cambridge University Press, 2002.

[2] L. Green, "Remote past and states in African-American English," *American Speech*, vol. 73, no. 2, pp. 115–138, 1998.

[3] J. R. Rickford, "Been in Black English," 1973, manuscript. University of Pennsylvania.

[4] ——, "Carrying the new wave into syntax: the case of Black English BIN," in *Variation in the form and use of language*, R. W. Fasold, Ed. Washington, DC: Georgetown University Press, 1975, pp. 98–119.

[5] L. Green, K. M. Yu, A. Neal, A. Whitmal, T. Powe, and D. Özyıldız, "Range in the use and realization of bin in African American English," *Language and Speech*, vol. 65, no. 4, pp. 958–1006, 2022.

[6] T. L. Weldon, *Middle-class African American English*. Cambridge University Press, 2021.

[7] N. Holliday, "Intonational variation, linguistic style, and the black/biracial experience," Ph.D. dissertation, New York University, New York, NY, 2016.

[8] E. E. Tarone, "Aspects of intonation in Black English," *American Speech*, vol. 48, no. 1/2, pp. 29–36, 1973.

[9] S.-A. Jun and C. Foreman, "Boundary tones and focus realization in African American English intonations," *The Journal of the Acoustical Society of America*, vol. 100, no. 4_Supplement, pp. 2826–2826, 1996.

[10] C. Mallinson and W. Wolfram, "Dialect accommodation in a biethnic mountain enclave community: More evidence on the development of African American English," *Language in Society*, vol. 31, no. 5, pp. 743–775, 2002.

[11] J. McLarty, "African American Language and European American English intonation variation over time in the American South," *American Speech: A Quarterly of Linguistic Usage*, vol. 93, no. 1, pp. 32–78, 2018.

[12] F. Cangemi, M. Krüger, and M. Grice, "Listener-specific perception of speaker-specific production in intonation," *Individual differences in speech production and perception*, pp. 123–145, 2015.

[13] N. Holliday, "Intonation and referee design phenomena in the narrative speech of black/biracial men," *Journal of English Linguistics*, vol. 49, no. 3, pp. 283–304, 2021.

[14] K. A. H. N. Arvaniti, A., "Variability, overlap and cue trading in intonation," To appear.

[15] W. E. Cooper and J. M. Sorensen, *Fundamental Frequency in Sentence Production*. New York, NY: Springer, 1981.

[16] A. Arnhold and A.-J. Kyröläinen, "Modelling the interplay of multiple cues in prosodic focus marking," 2017.

[17] G. Lohfink, A. Katsika, and A. Arvaniti, "Variability and category overlap in the realization of intonation," 2019.

[18] S. Gryllia, K. Marcoux, K. Jepson, and A. Arvaniti, "The many shapes of H*," May 2022, pp. 754–758.

[19] C. Kaland, "Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours," *Journal of the International Phonetic Association*, vol. 53, no. 1, pp. 159–188, Apr. 2023, publisher: Cambridge University Press.

[20] S.-E. Kim and S. Tilsen, "Planning for the future and reacting to the present: Proactive and reactive F0 adjustments in speech," *Journal of Phonetics*, vol. 104, p. 101322, May 2024.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[22] T. Kendall and C. Farrington, *The Corpus of Regional African American Language*, version 2023.06 ed. Eugene, OR: The Online Resources for African American Language Project, 2023. [Online]. Available: https://doi.org/10.7264/1ad5-6t35

[23] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2024, tex.date-added: 2008-04-28 08:04:52 -0700 tex.date-modified: 2022-07-15 14:57:13 -0400. [Online]. Available: http://www.praat.org/

[24] C. De Looze and S. Rauzy, "Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration," in *INTERSPEECH-2009*, 2009, pp. 2919–2922.

[25] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org/

[26] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[27] S. Moritz and T. Bartz-Beielstein, "imputeTS: Time Series Missing Value Imputation in R," *The R Journal*, vol. 9, no. 1, pp. 207–218, 2017.

[28] M. Gubian, F. Torreira, and L. Boves, "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts," *Journal of Phonetics*, vol. 49, pp. 16–40, Mar. 2015.

[29] A. Sardá-Espinosa, "Time-series clustering in R using the dtwclust package," *The R Journal*, 2019.

[30] A. Kassambara and F. Mundt, *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2020, r package version 1.0.7. [Online]. Available: https://CRAN.R-project.org/package=factoextra