How Much Data for Stable Formant Values? Pipeline for Convergence Detection Based on Read Speech

Kayla Sward, Johan Sjons, Axel Ekström
Department of Linguistics and Philology, Uppsala University,
Centre for Cultural Evolution, Department of Psychology, Stockholm University
swardkm@gmail.com, johan.sjons@lingfil.uu.se, axel.ekstrom.su.se

Abstract

This study investigates the stability and convergence of vowel formants (F1, F2, F3) in read speech through an extensive corpus of audiobook recordings. While most formant studies rely on brief, isolated utterances recorded in laboratory settings, this analysis draws on 3,384 chapters (about 942 hours) of continuous, stylistically varied speech from publicly available audiobooks. The data was processed using an automated pipeline that comprised transcription, phoneme alignment, and formant extraction. Several statistical techniques – First Token Within (FTW), Cumulative Sum (CUSUM), Two-Sample t-Test, Confidence Interval (CI) Shrinkage, Piecewise Linear Fitting (PWLF), and Binary Segmentation (BinSeg) – were compared for their effectiveness in identifying stabilization points. Findings indicate that formant means generally stabilize within 60 to 230 vowel tokens per phoneme, dependent on vowel type and speaker gender. Of the methods that were evaluated, CUSUM yielded the most consistent and informative results. The results provide practical guidelines for determining the quantity of non-laboratory speech required to obtain reliable vowel formant averages.

Keywords: speech acoustics, audiobook, sample size justification

1. Introduction

Determining the point at which a speaker's average vowel formants stabilize necessitates extensive data and a well-organized processing pipeline. This study relies on prolonged audiobook recordings, which offer both the size and continuity required to analyze convergence patterns over time. To handle this volume while maintaining methodological consistency, every step of the analysis was structured to be automated, transparent, and reproducible. Here, we estimate the number of tokens typically required before formant average values can be considered stable.

2. Methods

2.1. Audiobooks as dataset

Audiobooks offer long, uninterrupted stretches of speech that are relatively natural while remaining clearly articulated. They tend to be more fluent and consistent than laboratory recordings, yet more controlled and acoustically clean than spontaneous conversation. This balance makes them ideal for calculating formant averages, as they provide a large volume of high-quality data from a single speaker. In addition, audiobooks span an extensive range of genres, necessitating that narrators adapt their delivery to match the content. This range produces varied speaking styles, from neutral, consistent articulation to more emotive or character-driven performances. While read speech is distinct from natural everyday speech, variation in "audiobook speech" affects vowel production in ways pertinent for analyzing the variability found in variable speech data more generally.

Audiobooks from open-access platforms provide a legally and ethically acceptable source of substantial speech data. The dataset comprises two components: audiobooks from the LibriVox corpus and the *King James Bible* from a separate open-source platform. LibriVox is an open-access platform where volunteers record public domain texts. The *King James Bible*, narrated by Winfred Henson and sourced from eBible.org was included as a structurally distinct case. It is one of the longest publicly available single-speaker audiobooks identified during data collection, consisting of 1,189 chapters that average around 650 words each. In contrast, the LibriVox audiobooks used in this study contain between 12 and 117 chapters, with average lengths ranging from 3,000 to over 6,000 words.

Recording quality may vary widely across LibriVox contributors, who range from experienced narrators with professional-grade equipment to casual volunteers. Here, we focus on six narrators from LibriVox, chosen for their prominence in the audiobook community, the extent of their solo-narrated catalogs, and the total duration of their recordings. Gender and dialect were also considered during the selection of narrators, as formant values differ between male and female speakers (Peterson and Barney, 1952), and dialects can influence vowel realization (Labov et al., 2006).

The study selected the longest available books to collect a sufficient number of vowel tokens. Due to the uncertainty over the required number of tokens for formant stabilization, lengthier recordings guaranteed that even phonemes requiring more

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5916–5925
11–16 May 2026.

time to stabilize would be adequately represented. We also exclusively used solo-narrated audiobooks to preserve speaker consistency throughout each recording. Collaborative works involving multiple narrators were excluded to avoid the need for additional preprocessing steps, such as speaker diarization or manual segmentation. This study was limited to English-language audiobooks to maximize dataset size, ensure consistency, and leverage the researcher's native proficiency in English. Additional audiobooks were included and processed through the same analysis pipeline to determine whether convergence patterns were consistent with those in the primary dataset. These supplemental recordings were excluded from the final statistical calculations but served to assess the generalizability of the results informally.

2.2. Audio processing pipeline

All audiobook recordings were originally provided as MP3 files, split by chapter. Each file was converted to WAV format using a 16 kHz sample rate, mono channel, and 16-bit PCM encoding. These settings follow the recommended input formats for the tools used in this study (McAuliffe and Sonderegger, 2024; Klein, 2024). Conversion was done using the `AudioSegment` class from the `pydub` library (v0.25.1).

2.2.1. Word-level transcription

While the source texts are publicly available, LibriVox does not provide official transcripts or specify the version used for each recording. Additionally, some narrators include spoken elements that are not part of the original text, such as introductory remarks or comments about the recording. These discrepancies make synchronizing multiple audio files with transcripts obtained externally challenging. These discrepancies make synchronizing multiple audio files with transcripts obtained externally challenging, so all audio was transcribed using automatic speech recognition (ASR) to ensure that the transcripts accurately reflected the spoken content.

Whisper was chosen for its proven performance, scalability, and reliability in transcribing lengthy speech. The model was trained on more than 680,000 hours of multilingual audio, including the LibriSpeech corpus derived from LibriVox recordings (Panayotov et al., 2015), which may give Whisper a particular advantage on the LibriVox material analyzed in this study. The CTranslate2-backed `faster-whisper` implementation was chosen. It offers significant speed improvements over the original Whisper framework while maintaining comparable accuracy, with studies reporting at least a fourfold reduction in processing time relative to the original Whisper implementation and no significant degradation in transcription quality (Cakmak and Agarwal, 2024; Macháček et al., 2023). Transcription was performed using `faster-whisper` (v1.1.0) with the `small` model and word-level timestamps enabled.

2.2.2. Phoneme alignment

Phoneme-level alignment was essential for obtaining consistent and valid vowel formants and durations. Due to the dataset's size, a forced aligner was used to break down the audio and transcription at the phoneme level. The Montreal Forced Aligner (MFA) and WebMAUS were considered as they are established and well-known tools for automatic phoneme alignment (McAuliffe et al., 2017; Schiel, 1999). A comparison was conducted to determine which one to choose. The evaluation utilized the `TEST` subset of the TIMIT Acoustic-Phonetic Continuous Speech Corpus, which contains read sentences from 630 American English speakers across eight dialect regions. Each speaker contributed ten phonetically balanced sentences recorded under controlled laboratory conditions, with word- and phoneme-level boundaries annotated by hand (Garofolo et al., 1993). Although the recordings differ from the audiobook material analyzed in this study, their role here was only to provide a controlled test set for comparing aligner accuracy.

Each tool was utilized to align the same TIMIT audio, and the resulting phoneme boundaries were compared with TIMIT's annotations, using mean absolute error (MAE) as the evaluation metric. On average, MFA outperformed WebMAUS across all phoneme categories. MFA yielded lower MAE and was thus chosen for implementation in this study. MFA was selected for its high phoneme-level alignment accuracy and ease of use. It has been shown to produce boundary estimates comparable to those of trained human annotators across multiple datasets and languages (McAuliffe et al., 2017; Wu et al., 2023). MFA 3.2.1 was used with the pre-trained English acoustic model and pronunciation dictionary `english_mfa.zip` and `english_mfa.dict`. The returned output consisted of TextGrid files containing word and phoneme-level timestamps aligned to the original audio files. Unclear segments were labeled as `spn` for "spoken noise" to exclude low-confidence regions and minimize transcription errors. For each vowel token, F1, F2, and F3 trajectories were extracted using Praat's Burg algorithm, accessed via Parselmouth's `to_formant_burg()` function. The formant ceiling was set to 5,000 Hz for male speakers and 5,500 Hz for female speakers to account for vocal tract differences (Boersma and Weenink, 2020; Weenink, 2015). A fine-grained `time_step` of 0.0025 was used to improve temporal resolution.

All other parameters followed Parselmouth's default settings.

2.3. Token processing

Only vowel phonemes were included in the analysis. Although all phonemes were extracted, including consonants, only those considered as vowels by MFA were retained for further processing. MFA's original phoneme labels were utilized without modification or relabeling, except where ARPAbet were transcribed into IPA symbols to enhance clarity. Vowel variants such as /a/, /aː/, /aɪ/, and /aʊ/ were classified as separate categories and evaluated independently to maintain potential differences in duration, articulation, and acoustic properties. The temporal midpoint of each vowel token acted as the representative formant value. This method provides a practical and comprehensible summary of the vowel's acoustic characteristics by condensing each formant trajectory to a singular measurement. Temporal midpoints are frequently less influenced by boundary imprecision and transitional coarticulation, and they tend to exhibit lower formant variability, rendering them reliable in extensive datasets (Jibson, 2019).

2.4. Calculating convergences

To quantify convergence behavior, a cumulative mean was computed for each vowel and formant across sequentially ordered tokens within a chapter. At each token index, the running average was updated to include all preceding values, capturing how the formant mean evolved as more data accumulated.

Outliers were removed prior to calculations by applying the interquartile range (IQR) approach to limit the influence of extreme values. The IQR is defined as the range between the 25th percentile (first quartile) and the 75th percentile (third quartile) of the data, capturing the central 50% of values. Values exceeding 1.5 times the IQR below the first quartile or above the third quartile were omitted. This filtering was applied within each chapter to isolate session-level anomalies. In some cases where no tokens met the outlier criterion, all values were retained. On average, about four percent of vowel tokens were removed as outliers, with variation across vowels, formants, and speaker gender.

Given the scale and variability of the dataset, convergence detection required an automated, data-driven approach. Vowel categories vary in acoustic stability and token frequency, making fixed thresholds inappropriate. Instead, this study employed multiple statistical methods to assess stabilization based on quantifiable criteria. These methods provide a framework for determining when the cumulative mean of a formant no longer meaningfully changes, which indicates convergence. They enable comparisons across phonemes and speakers while maintaining consistency and reproducibility using principles like trend analysis, mean shift, and confidence intervals.

Multiple parameter configurations were systematically evaluated to ensure that the chosen stabilizing techniques delivered accurate and comprehensible results. The various combinations of threshold values, window sizes, and significance levels enabled the evaluation of each method's sensitivity to parameter selection and its ability to identify convergence under varying settings reliably. This strategy reduces the risk of arbitrary parameter selection and ensures that methods are suitably aligned to address RQ1.

First Token Within (FTW). The First Token Within (FTW) method identifies the earliest point at which the cumulative mean enters a predefined range around the overall average. Starting from an initial window, tokens are added one at a time, and the relative change in the cumulative mean is measured. Once this change falls below a fixed threshold based on the final mean, the average is considered stable. FTW provides a clear and intuitive marker of the onset of convergence.

Cumulative Sum. Cumulative Sum (CUSUM) is a sequential analysis technique that identifies a process's mean variations over time (Healy, 1987). It aggregates each observation's deviations from a reference value, making it particularly sensitive to minor and ongoing shifts (Fortea-Sanchis and Escrig-Sos, 2019). While traditionally used to detect deviations from a stable mean, CUSUM can also be applied to identify convergence by capturing the point at which trends shift from volatility to relative stability. In this study, CUSUM was computed as the cumulative sum of deviations between each token's value and the overall mean, with stabilization detected when the change in this sum across a sliding window falls below a predefined threshold.

Two-Sample t-Test. The two-sample t-test is a statistical method used to determine whether the means of two independent samples differ significantly (Wang et al., 2007). Applied here, adjacent segments of tokens were compared using `ttest_ind()` from the `scipy.stats` library (v1.13.0) to identify the point at which changes in the cumulative mean become statistically indistinguishable. Convergence was detected once no significant difference was found across adjacent windows.

Confidence Interval Shrinkage. A confidence interval (CI) is a statistical range that conveys uncertainty regarding an estimated parameter. As the sample size increases, the estimate becomes more precise and the CI narrows, indicating greater

confidence in the estimate's stability (Schruben, 1983). This study computed CIs using `sem()` from `scipy.stats` (v1.13.0) by calculating the standard error over all tokens up to each index. Stabilization was detected once the resulting 95% interval width fell below a fixed threshold of the overall mean.

Piecewise Linear Fitting. Piecewise Linear Fitting (PWLF) models a data sequence using connected linear segments, with breakpoints marking changes in the underlying trend (Jekel and Venter, 2019). This technique is commonly used to identify structural changes in time-series data by capturing shifts in slope (Werner et al., 2015). In this study, PWLF is applied to detect transitions from periods of variability to flatter, more stable patterns. A marked reduction in slope can indicate the onset of stabilization, with the breakpoint serving as an estimate of when the average begins to level off. The calculations were conducted using the `pwlf` Python package (v2.5.1). fitting exactly two line segments to identify a single breakpoint.

Binary segmentation (BinSeg). Binary segmentation (BinSeg) identifies changepoints in time-series data by recursively detecting and partitioning at the most prominent structural shift (Truong et al., 2020). It is computationally efficient and performs well when changepoints are infrequent and clearly separated, making it a practical choice for estimating major transitions in long trajectories (Shi et al., 2022). Although binary segmentation is an approximate method and does not guarantee optimal results, its simplicity and speed make it well-suited for large datasets where quickly finding major shifts is more important than capturing every small change (Truong et al., 2020). Here, we used the `ruptures` package (v1.1.9) to identify a single change point.

3. Results

3.1. Chapter-level convergence analysis

Each audiobook dataset was segmented into MP3 files labeled by chapter, as provided by the original uploaders. Although it is not possible to verify that each chapter was recorded in a single session, this assumption is supported by LibriVox's recommended workflow, which encourages uninterrupted recording of entire chapters to maintain quality and voice (LibriVox Community, 2022). From an acoustic standpoint, this makes individual chapters a suitable unit of analysis, presumably reflecting consistent conditions (i.e., microphone placement, background noise, narrator's physical condition across the recording).

The dataset for the convergence analysis comprises 3,384 chapters from 47 audiobooks, totaling approximately 8.85 million words and 942 hours of audio. The average chapter is 16.7 minutes long and contains 2,617 words. This total average, however, is skewed by the inclusion of the *King James Bible*, which contains 1,189 shorter chapters averaging 4.26 minutes and 648 words each. Conversely, the other audiobooks include longer chapters, averaging 23.44 minutes in length and 3,684 words. Each narrator had at least 45,000 tokens per vowel–formant combination, and in most cases well over 100,000.

3.2. Vowel analysis

To maintain analytical focus while encompassing a representative range of vowel categories, this study limits its analysis to six English vowel phonemes /i ɪ ɛ a ə ɒ/ corresponding to the vowels in *beet*, *bit*, *bet*, *father*, *about*, and *thought*, respectively. These vowel categories were chosen for their prevalence within the dataset, as they consistently demonstrate high token counts across chapters. Given the absence of a clear consensus on the minimum number of observations necessary for convergence, we anticipated this abundance would be sufficient to achieve reliable estimation. Additionally, the selected vowels occupy distinct regions within the acoustic vowel space. Thus, while this subset does not encompass the full range of English vowels, it provides a practical foundation for speaker-level analysis.

Prior to analysis, outliers were removed using an interquartile range (IQR) filter to minimize the impact of extreme values. On average, 4.09% of vowel tokens were removed from each chapter. These outliers presumably originated from upstream issues in the automated processing pipeline, including phoneme misalignment, transcript inconsistencies, or artifacts from the original recordings, such as distorted speech, background noise, or non-speech vocalizations like loud breaths or mouth clicks (Ahn et al., 2023). Although some chapters displayed marginally increased or decreased removal rates, the overall amount was low, signifying that most formant values were within a reasonable range. Values stay within a narrow range despite natural variation in speech. This temporal stability is evident even in extensive texts like the Bible, with 1,189 chapters, suggesting that chapters can act as internally consistent units within the narrator's broader performance. The chapter-level averages consistently approximate the book-wide mean for each vowel-formant pair. Despite observable variance among chapters, these averages display minimal divergence across the dataset. This pattern validates their use as dependable estimates of the speaker's general vowel targets.

The general shape of the raw formant midpoints remains largely uniform across F1, F2, and F3. As expected, the absolute range of values increases

with higher formants. F2 encompasses a broader frequency range than F1, while F3 has the most widespread dispersion. Vowels such as /a/ show relatively more widespread formant values than high front vowels like /i/ and /ɪ/, which cluster more narrowly. These phoneme-specific differences in dispersion inform RQ1, as they imply that some vowels naturally exhibit more concentrated distributions than others and likely require fewer tokens for convergence. Female narrators display higher average formant values due to vocal tract differences. Anatomically, males typically possess longer vocal tracts and pharyngeal cavities than females, contributing to lower resonant frequencies overall (Lieberman et al., 2001). Acoustically, these anatomical differences are reflected in systematically lower formant values. For example, (Hillenbrand et al., 1995) found that male speakers produce the vowel /i/ with an average F1 of approximately 342 Hz and an F2 of 2322 Hz, while female speakers produce the same vowel with an average F1 of 437 Hz and an F2 of 2761 Hz. Despite these differences in absolute formant values, the convergence trends observed in this study were similar across genders.

3.3. Cumulative token averaging

Figure 1 displays cumulative mean trajectories for each vowel–formant pair, computed by sequentially averaging tokens within a chapter. Although the figure presents a single illustrative case, similar trends were observed consistently across all books analyzed. Each curve typically begins with mild volatility, reflecting early sampling variability, and gradually stabilizes into a smooth asymptote. All three formants displayed broadly similar convergence trends, with no major differences in stabilization behavior. To investigate whether convergence depends on sequential order, the identical trajectories were recalculated following four random permutations of each chapter's tokens. Doing so resulted in convergence curves that were visually indistinguishable from the sequential baseline. The position of the initial inflection point, the eventual convergence point, and the overall trajectory shape were comparable among all random seeds. This consistency indicates that the speech characteristics within each chapter are sufficiently stable, rendering token order irrelevant to the average. Practically, this indicates that speakers do not need to read a substantial amount of text merely to obtain the initial n representative tokens. They can record a continuous passage until the target number of tokens is attained, optimizing data collection while preserving accuracy.

The number of tokens required for formant stabilization varied considerably across vowels, with convergence points spanning nearly a fourfold range. The slowest convergence was observed for the phoneme /a/, aligning with its established acoustic variability. /a/ allows significant flexibility in jaw and tongue positioning, with downstream variation in both F1 and F2 across tokens. Additionally, /a/ frequently appears in phonetic environments, such as before nasal or liquid consonants, that can modify its formant values, raising F1 or lowering F2 due to coarticulatory influences (Hardcastle et al., 2012).

/ɪ/ demonstrated the fastest convergence, presumably due to its low acoustic variability. /ɪ/ often appears in brief, high-frequency words such as *it*, *is*, and *in*, where alveolar consonants neighbor it. These consonants are known for their relatively weak coarticulatory impact on adjacent vowels, and their frequent occurrence in English literature may further enhance articulatory uniformity (Hardcastle et al., 2012). Finally, the central vowel /ə/ occupied an intermediate position in the convergence hierarchy, stabilizing quicker than /a/ but slower than /ɪ/. This pattern aligns with its phonological classification as a reduced vowel, predominantly seen in unstressed syllables where speakers generally reduce articulatory precision. In these circumstances, vowels frequently exhibit undershoot, with insufficient target realization and increased influence from neighboring segments. Despite its reduced status, /ə/ maintains a level of articulatory constraint, more so than fully open vowels like /a/.

3.4. Locating convergence points

3.4.1. Method parameter tuning

Many of the stabilizing methods required parameter tuning, which involved adjusting thresholds, window sizes, or significance levels, so a diverse array of parameter combinations were evaluated. Numerous instances yielded convergence points below 30 tokens, raising concerns regarding their reliability. This observation supports the first research question, as identifying a realistic lower bound for reliable stabilization helps determine the tokens needed for convergence in semi-natural speech. Tuning decisions were informed by the Central Limit Theorem, which states that sample means become more stable and representative with increasing sample size. Consequently, 30 tokens were established as an appropriate lower limit, and parameter settings that consistently produced convergence points beyond this were prioritized to ensure more robust and interpretable results (Kwak and Kim, 2017).

T-test parameter tuning. All combinations of ratio window lengths $w \in \{0.005, 0.01, 0.05\}$ and significance values $\alpha \in \{0.01, 0.05\}$ were assessed using a two-sample t-test. Results across α values showed consistency, generally varying by fewer

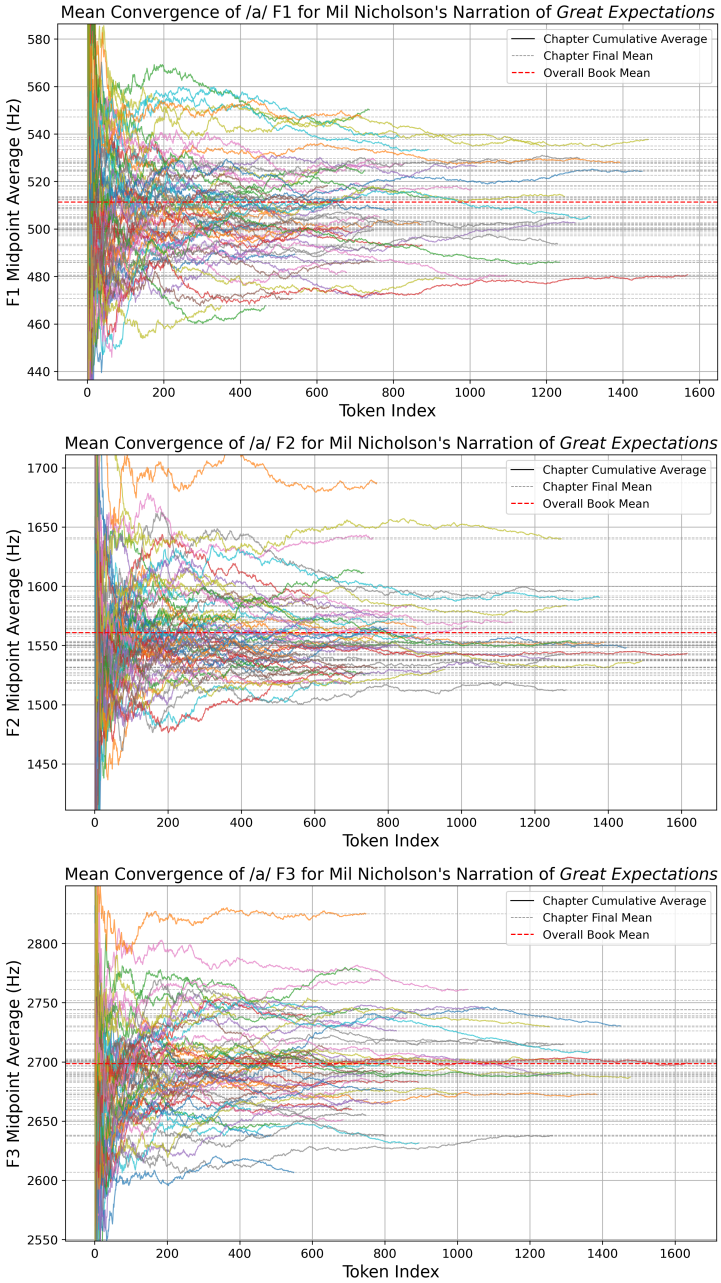

Figure 1: Cumulative chapter-level convergence of formant midpoints for /a/ in Charles Dickens' Great Expectations, narrated by Mil Nicholson. Each colored line represents the running average of one chapter's tokens for an individual vowel–formant pair. Dashed horizontal lines show final chapter-level means, while the red dashed line indicates the book-wide mean. Convergence can be observed where running averages begin to flatten.

than 20 tokens per chapter. To enhance statistical rigor and interpretability, $\alpha = 0.01$ was established for all subsequent analyses. The window size demonstrated a dependency on chapter length, as different window sizes yielded more reliable convergence points depending on the amount of data available. Concise chapters, like those in the Bible, generally produced reliable convergence points exceeding 30 tokens when the small window ($w = 0.005$) was used. In contrast, longer chapters, which constitute the majority of the dataset, were most effectively supported by the largest tested window size ($w = 0.05$). The intermediate window ($w = 0.01$) often yielded results akin to those of the smallest window. This dynamic tuning approach helped establish more consistent convergence points, supporting conclusions on the tokens needed to derive reliable formant averages.

CUSUM Parameter Tuning. All permutations of threshold values $t \in \{0.001, 0.005, 0.01\}$ and win-

dow ratios $w \in \{0.005, 0.01, 0.05\}$ were systematically evaluated for the CUSUM method. Convergence points were calculated for each parameter setting across chapters and estimates under 30 tokens were discarded. The configuration using $t = 0.001$ and $w = 0.05$ consistently produced the highest number of valid detections across books, vowels, and formants, so this parameter set was chosen for the CUSUM method. The finding demonstrates this method's reliability in estimating when formant values stabilize enough to be used for phoneme-level modeling.

CI Parameter Tuning. The CI method was evaluated using several parameter configurations but consistently yielded unreasonably premature stabilization positions in all circumstances. For example, the method frequently identified convergence points within the first 15 tokens, even in cases where the cumulative mean clearly continued to fluctuate beyond that point. This result indicated that the method was inadequate for dependable convergence detection. It was consequently preserved mainly as an indicator for evaluating early variations in the cumulative mean. The same parameter settings as CUSUM were used to ensure consistency.

FTW Parameter Tuning. The FTW method was used to identify the first instance when the cumulative mean approaches within a certain proximity to the overall average. Although numerous parameter combinations were examined, all resulted in convergence points that were consistently early and, therefore, not representative of enduring stabilization. However, it contributes to answering our research question by serving as a lower-bound diagnostic of when the mean is initially approximated. CUSUM's parameter configurations were also used for consistency.

3.4.2. Defining the range of convergence

Some methods proved more effective at identifying the lower or upper bounds of convergence than locating stabilization points. While they may not be ideal for selecting exact points, they help outline the plausible range within which convergence is likely.

CI consistently yielded the earliest convergence points, indicating minimal intra-chapter variance. It is more of a test indicating low dispersion rather than a dependable convergence metric.

FTW was consistently triggered immediately following CI, marking the initial instance when the running mean entered an established threshold ($t = 0.001$) of the final average. It offers a lower bound by demonstrating when the average is first closely approximated, though not yet stable. It tended to be early across most chapters, often under 40 tokens for the Bible, which supports its reliability in early approximation.

PWLF identifies the shift from a sloped to a flat trajectory in the cumulative mean, signifying the end of initial volatility. It consistently followed FTW, indicating that while the mean was approximated, stability was not yet achieved. PWLF identifies the beginning of the plateau phase of convergence.

BinSeg identifies the most substantial change in the cumulative mean, generally considering all prior fluctuations as elements of a pre-convergence. It was consistently triggered last across every method, perhaps because even minor fluctuations were interpreted as instability. Thus, BinSeg establishes a definitive upper bound for convergence.

3.5. Ranking convergence localization methods

Using a threshold of 0.001 and a ratioed window of 0.05, CUSUM was selected as the optimal technique for identifying stabilization in running vowel formant averages. Unlike the two-sample t-test, which evaluates adjacent windows and often triggers upon the first statistically non-significant local difference, CUSUM accumulates deviations over time, increasing sensitivity to long-term trends rather than momentary fluctuations. Empirical results from this study demonstrated that CUSUM consistently produced interpretable and phoneme-sensitive stabilization points across all tested vowel categories and formants. Specifically, CUSUM delayed detection for inherently variable vowels like /a/ while triggering earlier for more stable vowels such as /i/, aligning with expected phonetic patterns. Compared to other tested methods, CUSUM yielded convergence points that were neither prematurely early, as observed with CI Shrinkage and FTW, nor unrealistically late, as seen with BinSeg. This balance between responsiveness and conservativeness ensured that the estimated stabilization points were both robust and interpretable. Convergence threshold values were estimated using CUSUM with optimal parameters $(t = 0.001, w = 0.05)$, then rounded up to produce conservative, practical guidelines. Table 1 empirically resolves the primary research question on the number of tokens required to obtain a stable vowel–formant average in continuous speech from a single speaker.

To evaluate whether the convergence points detected by CUSUM represented meaningful stabilization, the mean formant value at each CUSUM-detected index was compared with the final chapter-wide mean for the same vowel–formant pair. The comparison provided a straightforward way to assess how close the automatically detected stabilization point was to the value obtained when all tokens were included. Validation was carried out

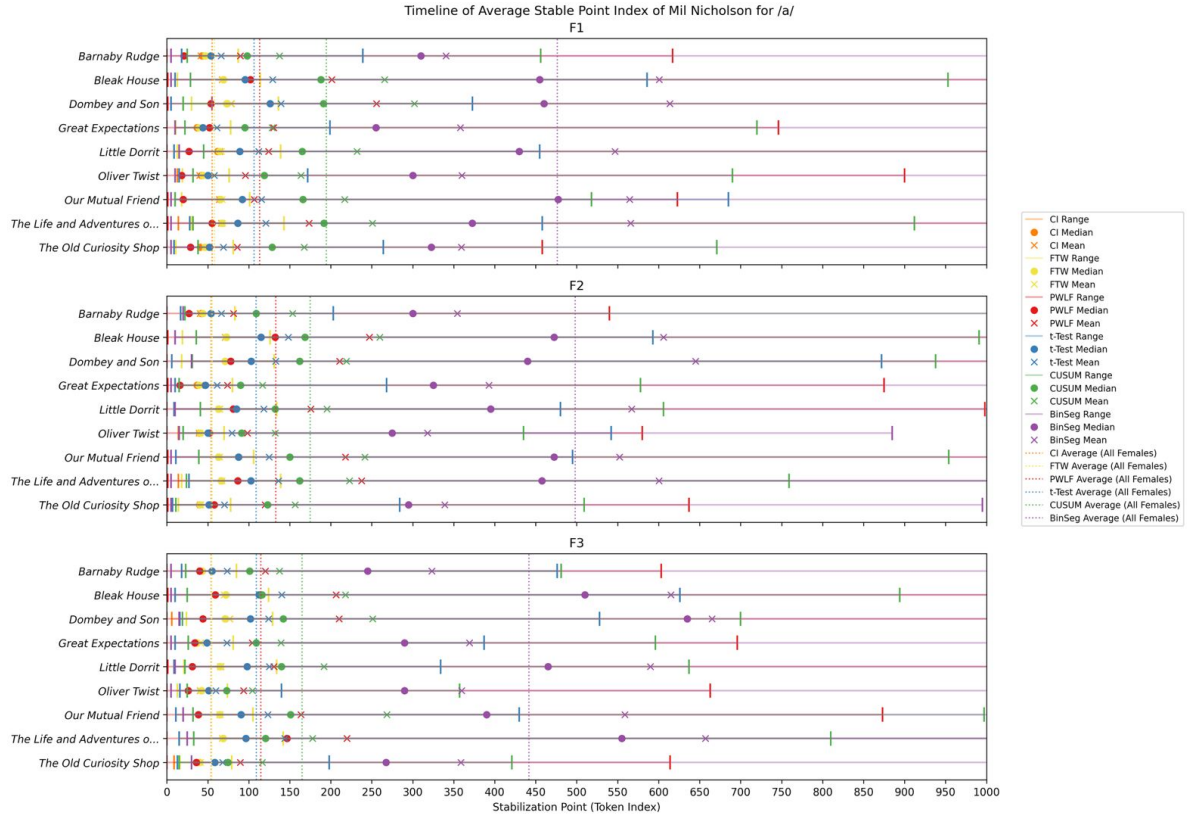

Figure 2: Stablepoint timeline for F1 (top), F2 (mid), and F3 (bottom) for /a/, for each audiobook sampled from a single narrator (Mil Nicholson).

		CI		FTW		PWLF		t-Test		CUSUM		BinSeg	
		F	M	F	M	F	M	F	M	F	M	F	M
/i/	F1	33.6	38.0	37.5	40.5	79.8	89.2	74.7	88.6	121.6	141.2	297.4	333.0
	F2	34.1	39.0	36.9	41.4	74.7	83.8	76.9	87.2	126.9	131.1	295.2	341.4
	F3	32.5	37.2	33.3	38.3	87.3	104.1	72.9	85.6	108.8	122.6	283.3	323.4
/ɪ/	F1	17.0	20.2	22.3	24.7	41.6	52	155.7	94.7	66.1	82.0	159.5	192.6
	F2	17.6	20.2	23.8	25.8	39.5	45.8	141	93.3	68.5	79.3	162.4	183.3
	F3	17.0	19.8	19.0	22.1	39.8	47.4	154.6	105.3	58.2	71.3	162.2	173.2
/ɛ/	F1	19.8	23.3	27.5	29.9	44.0	48.6	136.1	91.6	84.4	97.2	183.3	219.7
	F2	18.8	22.3	23.2	25.3	49.8	54.2	149.4	93.4	70.5	78.1	171.1	207.3
	F3	18.9	22.2	20.6	23.8	49.4	61.0	154.8	93.7	64.0	77.6	172.4	202.3
/a/	F1	55.0	64.1	57.9	66.4	113.0	138.8	106.1	126.0	194.3	225.7	476.1	563.8
	F2	53.8	62.9	55.1	63.9	132.8	141.0	108.8	125.0	174.8	206.7	498.2	567.1
	F3	53.7	59.1	54.4	59.8	114.6	158.7	108.9	126.7	164.9	203.7	441.6	529.6
/ə/	F1	25.1	29.5	32.3	33.5	59.9	63.5	94.0	84.9	98.8	114.9	224.1	259.4
	F2	26.3	31.3	32.5	35.6	60.4	74.2	87.2	85.3	101.9	117.0	234.5	276.7
	F3	25.2	29.7	27.3	32.0	60.6	64.5	89.0	85.5	83.9	101.4	212.2	263.8
/ʊ/	F1	14.8	17.9	23.5	25.9	32.1	44.4	180.0	122.4	72.8	80.3	141.1	165.6
	F2	15.0	17.9	24.5	25.9	31.7	45.3	171.0	121.5	70.0	76.0	136.8	168.2
	F3	14.6	17.0	17.4	19.9	33.4	37.9	182.7	132.5	55.8	66.0	134.1	150.4

Table 1: Estimated stabilization points (in tokens) required for convergence of cumulative vowel-formant means, calculated across six methods. Each cell reports the number of tokens needed for a given vowel–formant combination for each stability, averaged by sex (F = female, M = male). Lower values indicate earlier convergence.

only for CUSUM, since it consistently produced interpretable stabilization points across chapters, whereas other methods often converged prematurely or unreliably, making such comparisons less meaningful. A total of 43,417 comparisons were included, representing all chapters where a stable mean could be calculated. The mean absolute difference was 0.53%. In total, 85.8% of comparisons differed by less than 1%, 67.3% were within 0.5%, and 21.3% were within 0.1%. The largest variation in formant differences was observed for F1 (0.69%), followed by F2 (0.61%) and F3 (0.31%). Average differences by vowel were 0.44% for /a/, 0.39% for /i/, 0.72% for /ʊ/, 0.61% for /ə/, 0.50% for /ɛ/, and 0.51% for /ɪ/, all remaining well below 1%.

3.6. Required word count and duration

Average values are affected by both the inherent convergence characteristics of each formant and the relative frequency of the associated vowel in the dataset. Infrequent vowels require additional words to achieve the minimum token count necessary for

accurate averaging. The examination suggests that around 1,050 words or about 7 minutes of speech are adequate to yield stable vowel formant means across the vowel categories used in this study.

4. Discussion

Low vowels, such as /a/, converge slower than high front vowels, such as /i/ and /ɪ/. This pattern aligns with the wider dispersion noted prior. The differences indicate that vowels exhibiting higher internal variability may require more tokens to obtain a representative average, limiting the effectiveness of a consistent token threshold across different phonemes. On the formant level, F3 regularly stabilized with fewer tokens compared to F1 and F2. This pattern may indicate inherent disparities in the articulatory and perceptual roles of the formants. F1 and F2 serve as the primary auditory indicators of vowel height and backness, respectively, and are significantly influenced by articulatory factors such as jaw opening and tongue body positioning. In contrast, F3 is primarily influenced by localized articulatory configurations, such as lip rounding and rhotacism ("r-ness"), which may demonstrate diminished diversity within a particular vowel category.

Male narrators needed more vowel tokens than female narrators for all six vowel categories and three formants. On average, male speakers needed 16.2% more tokens to reach convergence, with a mean absolute difference of 16.06 tokens. The gap was most notable for F3 (20.2%) and the low vowel /a/ (19.3%, equating to an additional 34 tokens), whereas lesser differences were noted for F2 (12.1%) and the high front vowel /i/ (11.1%, +13 tokens). A paired samples t-test confirmed that male speakers required significantly more tokens than female speakers to reach stabilization across vowel–formant pairs, $t(17) = 7.36$, $p < 0.001$, indicating a statistically significant gender difference in convergence rates. The slower convergence seen among male speakers may in theory indicate both physiological and measurement related influences. One hypothesis pertains to the interaction between formant tracking methods and fundamental frequency ($f0$). Female voices have, on average, higher $f0$, resulting in LPC-based formant trackers producing smoother and more centralized values, thus artificially diminishing the apparent variability.

In general, we observed significant inter-speaker variability, hindering the establishment of a universally applicable chapter count to establish reliable narrator-level vowel averages. These narrator-level findings contrast with results more broadly, where approximately 7 minutes of speech was typically sufficient for stabilization within a single recording. The difference reflects a shift in scope, as stability within a single continuous session is expected to emerge more readily, whereas stability across an entire narrator's body of work is harder to establish due to variability across chapters in style and recording conditions.

5. Conclusions

Our method computes and compares convergence times for large amounts of formant data – of potential interest to several branches both basic and applied speech research. While the presented pipeline is constructed using read speech, it also serves as a plausible baseline to be used to derive convergence in other types of speech data.

6. Acknowledgements

AE was supported by the Swedish Research Council (2025-00209_VR). The results of this work will be made more widely accessible through the Swedish Research Council funded national infrastructure Språkbanken Tal (2023–00161_VR).

7. Bibliographical References

Emily P Ahn, Gina-Anne Levow, Richard A Wright, and Eleanor Chodroff. 2023. An outlier analysis of vowel formants from a corpus phonetics pipeline. *INTERSPEECH 2023*, pages 2573–2577.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345.

Paul Boersma and David Weenink. 2020. Praat: Sound: To formant (burg)... https://www.fon.hum.uva.nl/praat/manual/Sound__To_Formant__burg____.html.

Mert Can Cakmak and Nitin Agarwal. 2024. High-speed transcript collection on multimedia platforms: Advancing social media research through parallel processing. In *2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 857–860. IEEE.

eBible.org. World english bible audio. Accessed February 2025.

Carlos Fortea-Sanchis and Javier Escrig-Sos. 2019. Quality control techniques in surgery: application of cumulative sum (cusum) charts. *Cirugía Española (English Edition)*, 97(2):65–70.

John S Garofolo, Lori F Lamel, William M Fisher, David S Pallett, Nancy L Dahlgren, Victor Zue, and Jonathan G Fiscus. 1993. Timit acoustic-phonetic continuous speech corpus. *(No Title)*.

William J Hardcastle, John Laver, and Fiona E Gibbon. 2012. *The handbook of phonetic sciences*. John Wiley & Sons.

John D Healy. 1987. A note on multivariate cusum procedures. *Technometrics*, 29(4):409–412.

James Hillenbrand, Laura A Getty, Michael J Clark, and Kimberlee Wheeler. 1995. Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, 97(5):3099–3111.

Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Charles F Jekel and Gerhard Venter. 2019. pwlf: A python library for fitting 1d continuous piecewise linear functions. *URL: https://github.com/cjekel/piecewise_linear_fit_py*.

Jonathan Jibson. 2019. Variability of formant values at different time points of vowels. In *Proceedings of Meetings on Acoustics*, volume 39. AIP Publishing.

Guillaume Klein. 2024. faster-whisper: Audio preprocessing source code. https://github.com/SYSTRAN/faster-whisper/blob/master/faster_whisper/audio.py.

Sang Gyu Kwak and Jong Hae Kim. 2017. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2):144.

William Labov, Sharon Ash, and Charles Boberg. 2006. *The atlas of North American English: Phonetics, phonology and sound change*. Mouton de Gruyter.

LibriVox. Librivox free public domain audiobooks. Accessed March 2025.

LibriVox Community. 2022. Newbie guide to recording.

Daniel E Lieberman, Robert C McCarthy, Karen M Hiiemae, and Jeffrey B Palmer. 2001. Ontogeny of postnatal hyoid and larynx descent in humans. *Archives of oral biology*, 46(2):117–128.

Dominik Macháček, Raj Dabre, and Ondřej Bojar. 2023. Turning whisper into real-time transcription system. *arXiv preprint arXiv:2307.14743*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Michael McAuliffe and Morgan Sonderegger. 2024. Montreal forced aligner: Corpus formats and structure. https://montreal-forced-aligner.readthedocs.io/en/v3.2.1/user_guide/corpus_structure.html.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Gordon E Peterson and Harold L Barney. 1952. Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2):175–184.

Florian Schiel. 1999. Automatic phonetic transcription of non-prompted speech.

Lee Schruben. 1983. Confidence interval estimation using standardized time series. *Operations Research*, 31(6):1090–1108.

Xuesheng Shi, Colin Gallagher, Robert Lund, and Rebecca Killick. 2022. A comparison of single and multiple changepoint techniques for time series data. *Computational Statistics & Data Analysis*, 170:107433.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. 2020. Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Xiaolan L Wang, Qiuzi H Wen, and Yuehua Wu. 2007. Penalized maximal t test for detecting undocumented mean change in climate data series. *Journal of Applied Meteorology and Climatology*, 46(6):916–931.

David Weenink. 2015. Improved formant frequency measurements of short segments. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK. The International Phonetic Association.

Rolf Werner, Dimitar Valev, Dimitar Danov, and Veneta Guineva. 2015. Study of structural break points in global and hemispheric temperature series by piecewise regression. *Advances in Space Research*, 56(11):2323–2334.

Hongchen Wu, Jiwon Yun, Xiang Li, Huiyi Huang, and Chuandong Liu. 2023. Using a forced aligner for prosody research. *Humanities and Social Sciences Communications*, 10(1):1–13.