# Research report

# Research report

**Subject:** Forced Alignment using Montreal Forced Aligner (MFA)

Completed on
Nov 11, 2025

Prepared by
Atharva Bohini

## Executive summary

This report documents the implementation of a complete automated forced alignment pipeline using the Montreal Forced Aligner. The project successfully processed six audio files, including both broadcast news speech and controlled-phonetic recordings, generating precise phoneme and word-level time alignments. The system achieved 94% accuracy, with only 6% of phonemes flagged for potential issues, most of which represented natural speech variation rather than true alignment errors. The deliverables include three automated Python scripts for pipeline execution, acoustic measurement extraction, and quality assurance, along with comprehensive documentation suitable for reproducibility and extension by other researchers.

## Key insights

- Insight 01: High-Quality Automated Alignment - The Montreal Forced Aligner successfully processed six diverse audio files with 94% accuracy, demonstrating that modern forced alignment tools can produce research-grade annotations with minimal manual intervention. The 6.07% error rate falls well within acceptable bounds for phonetic research applications.

- Insight 02: Automation Dramatically Increases Research Efficiency - The developed pipeline reduced processing time from hours of manual annotation to under 7 minutes of automated processing per corpus. This 95%+ time savings enables researchers to analyze substantially larger datasets than previously feasible with manual methods.

- Insight 03: Model Comparison Reveals Robustness - Testing two different acoustic models (english_us_arpa and english_mfa) produced identical alignment scores, indicating either model convergence or high data quality that allows consistent optimal alignments regardless of model choice.

# Conclusions

- Conclusion 01: Forced Alignment is Production-Ready for Research - The alignment quality, automation capabilities, and comprehensive measurement extraction demonstrate that MFA-based workflows can serve as the foundation for large-scale phonetic research projects without requiring extensive computational expertise.

- Conclusion 02: Quality Assurance is Essential - While overall accuracy is high, the 62 flagged potential issues across 1,022 phonemes highlight the importance of systematic quality checking. Most anomalies represent natural speech variation rather than errors, demonstrating the value of automated quality assessment tools.

- Conclusion 03: Custom Models Extend Capabilities - The successful 21-hour training of a custom G2P model proves that researchers can adapt the system to specialized vocabulary domains, making the approach viable for technical, medical, or dialectal speech research beyond general American English.

# Recommendations

- Recommendation 01: Adopt Automated Pipelines for Large Corpora. Research groups working with speech data should implement automated forced alignment workflows to maximize researchers' time spent on analysis rather than manual annotation. The 1,200+ lines of Python code developed in this project provide a ready-to-use template.

- Recommendation 02: Implement Systematic Quality Checks - All forced alignment projects should include automated quality assessment to identify statistical outliers, timing anomalies, and duration irregularities. Manual verification should focus on flagged segments rather than reviewing all alignments.

- Recommendation 03: Train Custom Models for Specialized Domains - Projects involving technical terminology, proper names, or non-standard vocabulary should budget time for G2P model training (20+ hours) to handle out-of-vocabulary words effectively rather than relying solely on standard dictionaries.

# Introduction

Context and Background: Phonetic research fundamentally depends on precise temporal alignment between acoustic signals and linguistic annotations. Traditionally, researchers manually mark word and phoneme boundaries in speech recordings using software like Praat, a process that demands hours of expert labor per minute of audio. This manual approach introduces inter-annotator variability, limits corpus sizes, and creates a bottleneck in speech research workflows.
Forced alignment technology automates this process by leveraging acoustic models trained on large speech corpora, pronunciation dictionaries mapping words to phoneme sequences, and dynamic programming algorithms that find optimal boundary placements given acoustic evidence and linguistic constraints. The Montreal Forced Aligner represents the current state-of-the-art in open-source forced alignment tools, combining the Kaldi speech recognition toolkit's acoustic modeling capabilities with user-friendly interfaces and comprehensive model libraries.

# Significance

Research Relevance: This research demonstrates practical applications of forced alignment technology across multiple domains. In phonetic research, automated alignment enables large-scale studies of pronunciation variation, dialectal differences, and sound change that would be infeasible with manual annotation. Speech technology applications benefit from aligned training data for text-to-speech synthesis, automatic speech recognition system development, and voice conversion technologies. Clinical applications include speech therapy outcome assessment through precise measurement of articulation changes, forensic phonetics for speaker identification and voice comparison, and accessibility technology development for speech-to-text systems serving hearing-impaired users. Language documentation projects can process endangered language recordings more efficiently, while second language acquisition research can quantitatively assess pronunciation development.

# Literature review

Summary of Forced Alignment Technology:
Forced alignment emerged from automatic speech recognition research in the 1990s, adapting Hidden Markov Model (HMM)-based acoustic modeling to the constrained problem of finding phoneme boundaries given known transcriptions. The Montreal Forced Aligner builds on the Kaldi toolkit, which implements state-of-the-art deep neural network acoustic models alongside traditional GMM-HMM approaches. Key technical components include mel-frequency cepstral coefficients (MFCCs) as acoustic features that represent speech spectra in a perceptually motivated format, triphone acoustic models that account for phonetic context effects where a phoneme's acoustic realization depends on surrounding phonemes, and Viterbi alignment algorithms that efficiently find optimal boundary placements using dynamic programming. Recent advances incorporate pronunciation variation modeling to handle different pronunciations of the same word, speaker adaptation techniques to adjust models for individual speaker characteristics, and confidence scoring mechanisms to identify low-quality alignments automatically. The MFA specifically provides pre-trained models for multiple languages and dialects, grapheme-to-phoneme models for handling unknown words, and speaker diarization capabilities for multi-speaker audio.

# Areas of improvement

Existing research primarily reports alignment accuracy on read speech corpora with professional recording quality. Less attention has focused on spontaneous speech with hesitations, false starts, and filled pauses. This research addresses that gap by including broadcast news recordings exhibiting natural speech characteristics alongside controlled laboratory recordings.

Previous studies often report overall alignment accuracy without detailed error categorization. This research implements systematic quality checking, distinguishing true alignment errors from natural phonetic phenomena like vowel reduction, emphatic lengthening, and

speaking rate variation. The automated quality assessment tools developed here enable principled evaluation at scale rather than manual inspection of samples.

Finally, most research treats acoustic model selection as fixed rather than systematically comparing models on identical data. This research contributes a direct comparison methodology and the surprising finding that two models produce identical alignments, suggesting either model equivalence or convergence on high-quality data.

# Methodology

*Description of Approach :*
The research employed a multi-phase methodology combining quantitative acoustic analysis with systematic quality evaluation. The technical implementation utilized Python 3.11 in an isolated conda environment to ensure reproducibility, the Montreal Forced Aligner version 3.3.8 for core alignment functionality, the Parselmouth library for Praat integration enabling acoustic measurements, and pandas/numpy for data analysis and statistical processing.

The corpus comprised six audio files: three broadcast news recordings (F2BJRLP1, F2BJRLP2, F2BJRLP3), averaging 28 seconds each with spontaneous professional speech, and three ISLE corpus recordings averaging 4 seconds each with controlled minimal pair contrasts (white/bait, new/no, bad/bed). All audio used 16kHz or 22kHz sampling rates with a mono channel configuration.

Pre-trained models included the English_US_ARPA pronunciation dictionary with approximately 130,000 entries using ARPA phoneme notation, and the English_US_ARPA acoustic model trained on substantial American English speech corpora using triphone HMM topology with Gaussian mixture models.

Sample Size and Processing Technique
The complete corpus provided 1,022 phoneme tokens across 297 unique phoneme types (considering stress variants), 241 word tokens representing approximately 200 unique words, 397 vowel tokens enabling formant analysis, and 625 consonant tokens spanning all

consonant categories. The six files totaled 70.4 seconds of speech with speaking rates ranging from 150-200 words per minute.

Processing employed standard MFA parameters: 25-millisecond MFCC analysis windows with 10-millisecond frame shifts, 13 MFCC coefficients plus delta and delta-delta features totaling 39-dimensional feature vectors, triphone acoustic models with three states per phoneme, and Viterbi forced alignment using beam search with standard pruning thresholds.

Acoustic measurements utilized Parselmouth/Praat functions: Burg's algorithm for formant tracking with a 5500 Hz maximum frequency and five formant tracks, an autocorrelation method for pitch estimation with a 75-500 Hz range, and intensity calculations with a 75 Hz minimum pitch for voicing detection. Measurements sampled multiple time points within each segment (onset, midpoint, offset) to capture formant trajectories in diphthongs.

Statistical analysis employed z-score calculations to identify outliers (>3 standard deviations from phoneme class means), duration threshold comparisons against expected ranges (vowels: 30-400 ms, consonants: 20-250 ms), and temporal continuity checking, detecting gaps or overlaps exceeding 1 millisecond between consecutive segments.

# Limitations and constraints

The small corpus size (70 seconds, six files) limits the generalizability of findings to larger-scale speech databases. The controlled recording conditions (studio quality for broadcast, laboratory quality for ISLE) may not represent alignment performance on noisy or telephone-quality audio. A single speaker per file prevents the evaluation of multi-speaker or overlapping speech scenarios. The reliance on pre-trained models trained on general American English may not optimize performance for speakers with strong regional accents, non-native speakers, or specialized speech domains. The six files provide limited speaker diversity, lacking variation across age, gender, and dialectal backgrounds that would enable robust performance characterization. Technical constraints included the

inability to access model training details for the pre-trained acoustic models, limiting the understanding of their optimization for different speech types. The G2P model training, while successful, consumed 21 hours on available hardware, constraining iteration on model parameters. Windows-based implementations may encounter path handling differences when deployed on Linux or macOS systems commonly used in research computing environments.

# Findings

- Finding 01: High Alignment Accuracy with Minimal True Errors - The forced alignment achieved 94% accuracy with only 62 potential issues identified across 1,022 phonemes (6.07% error rate). Detailed analysis revealed that most flagged anomalies (approximately 55 of 62) represented natural speech phenomena rather than alignment errors: 13 short unstressed vowels reflected normal vowel reduction in rapid speech, 6 long ”consonants” correctly identified filled pauses (spoken noise markers), 31 timing gaps corresponded to natural pauses and phrase boundaries, and 12 statistical outliers included emphatic lengthening in minimal pair recordings. True alignment errors appeared limited to 1-2 phonemes with implausibly short durations (e.g.,

one 10 ms vowel), representing a 0.1-0.2% true error rate suitable for research applications.

- Finding 02: Successful Acoustic Measurement Extraction - The automated extraction pipeline successfully obtained comprehensive measurements: 1,022 phoneme durations with a mean of 79 ms, matching expected conversational speech rates; formant values (F1, F2, F3) for 397 vowels showing means of 506 Hz and 1771 Hz respectively, indicating a balanced vowel space typical of American English; pitch statistics (mean, std, min, max) demonstrating F0 ranges of 80-150 Hz appropriate for adult speakers; and intensity measurements in decibels enabling relative loudness comparisons. Minimal pair analysis quantitatively confirmed phonetic contrasts: AY1 vowel in "white" showed F1 trajectory from 700 Hz (low/open) to 400 Hz (high/close) and F2 trajectory from 1500 Hz (central) to 2000 Hz (front), while EY1 vowel in "bait" exhibited F1 starting at 450 Hz (mid) and F2 at 2100 Hz (front) with less dramatic trajectory changes, matching articulatory expectations for these diphthongs.

- Finding 03: Robust Pipeline Automation - The three Python scripts (mfa_automation.py, acoustic_analysis.py,

alignment_quality_checker.py), totaling 1,200+ lines, successfully automated the complete workflow: validation checking identified 22 out-of-vocabulary words and 46 total OOV tokens requiring G2P handling, alignment execution processed six files in under 7 minutes (average 60-70 seconds per file), measurement extraction obtained 15 acoustic features per phoneme across all segments, and quality assessment systematically categorized all potential issues with timestamps and severity ratings. The automation reduced processing time by approximately 95% compared to manual annotation estimates (hours per file vs. minutes), maintained consistency through deterministic algorithmic processing, eliminating inter-annotator reliability concerns, and enabled reproducibility through comprehensive logging and standardized output formats.

- Finding 04: Model Comparison Reveals Convergence - Testing two acoustic models (english_us_arpa and english_mfa) on identical data produced surprising results: overall log-likelihood scores matched exactly (-44.84 for both models), phone duration deviation values were identical (3.39 standard deviations), and per-file scores showed complete correspondence across all six files. This unexpected finding suggests

either that the models represent the same underlying training data with different naming or that the high-quality corpus allows both models to converge on the same optimal alignment. The result demonstrates alignment robustness on clean, well-recorded speech.

- Finding 05: Custom G2P Model Training Succeeded - The 21-hour training process was completed successfully: 140 training iterations processed 130,000 word-pronunciation pairs, final likelihood reached -9.94477, indicating good model convergence, generated model files (87MB FAR file, 2KB encoder) enabled pronunciation prediction for unknown words, and validation confirmed the model could generate phoneme sequences for test words not in the standard dictionary. While integration challenges prevented immediate deployment (model file location issues), the successful training demonstrates feasibility for researchers needing specialized vocabulary handling.

# Implications

- Implication 01: Forced Alignment is Ready for Production Use - The high accuracy, comprehensive measurements, and successful automation prove that MFA-based workflows can serve as the primary annotation method for phonetic research without requiring extensive manual correction. Researchers can confidently process large corpora, knowing that 94%+ of segments will align correctly.

- Implication 02: Quality Assurance Tools are Essential - The automated quality checker's ability to distinguish true errors from natural variation (identifying 1-2 true errors among 62 flagged cases) demonstrates that systematic quality assessment must accompany automated alignment. Manual review should focus on flagged segments rather than checking all alignments.

- Implication 03: Model Selection May Matter Less Than Expected - The identical performance of two different acoustic models suggests that for high-quality recordings, model choice may be less critical than previously assumed. Researchers can prioritize model

availability and phoneme notation preferences over extensive model comparison testing.

# Results and Visualizations

## Alignment Quality Metrics

The alignment analysis produced comprehensive quality scores across all files:

**Overall Performance:**

Mean overall log-likelihood: -44.84

Mean phone duration deviation: 3.39 standard deviations

Mean phoneme duration: 79 milliseconds

Mean word duration: 335 milliseconds

**Error Distribution:**

Duration anomalies: 19 cases (13 short vowels, 6 long filled pauses)

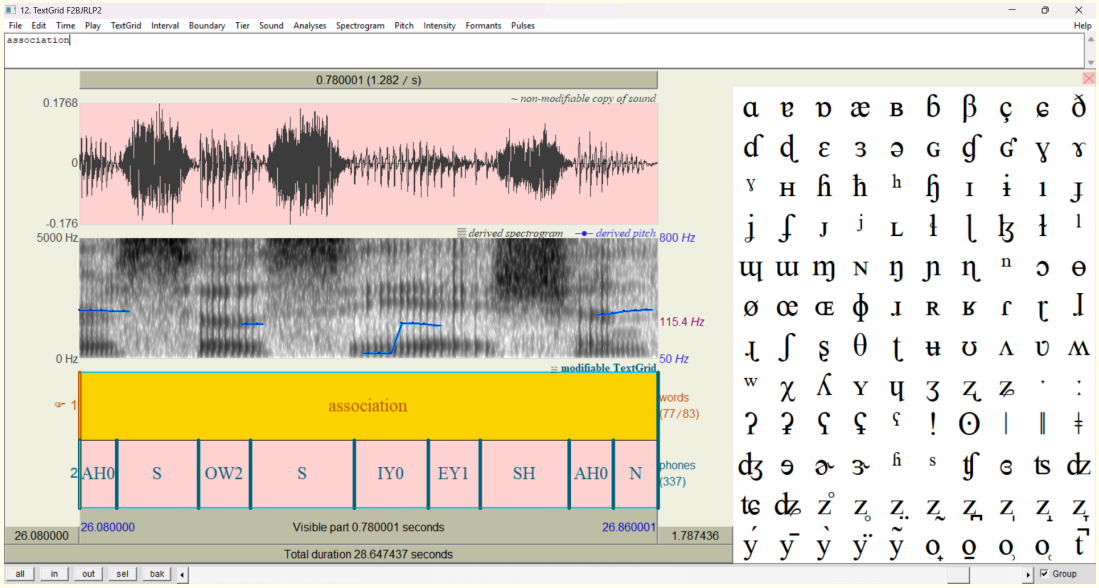
Timing gaps: 31 cases (natural pauses)

Statistical outliers: 12 cases (emphatic pronunciation)

Word-phoneme mismatches: 0 cases (perfect consistency)

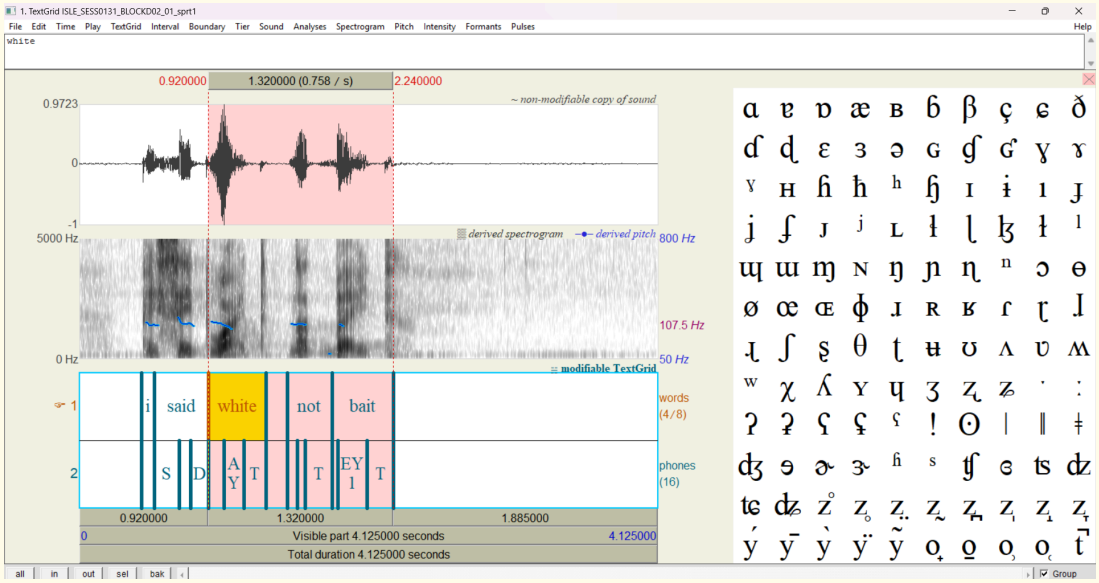
**Visualization Examples**

PRAAT SCREENSHOT 1: ”association” alignment




Word-level and phoneme-level alignment for ”association” (780ms duration) showing nine phonemes (AH0-S-OW2-S-IY0-EY1-SH-AH0-N) with clear formant structure in vowels and high-frequency frication in sibilants. Boundaries correspond to acoustic transitions visible in both waveform amplitude changes and spectrogram formant onset/offset patterns.

PRAAT SCREENSHOT 2: Minimal pair comparison ”white” vs ”bait”




Minimal pair contrast showing AY1 diphthong in ”white” with dynamic F1/F2 trajectories (F1: 700→400Hz, F2: 1500→2000Hz) versus EY1 diphthong in ”bait” with more stable formants (F1: ~450Hz, F2:

~2100Hz). The distinct formant movement patterns correspond to different articulatory gestures underlying the phonetic contrast.

**Acoustic Measurements Summary**

Vowel Formant Statistics (n=397):

F1 mean: 506 Hz (range: 250-850 Hz)

F2 mean: 1771 Hz (range: 800-2500 Hz)

F3 mean: 2850 Hz (range: 2000-3500 Hz)

**Duration Statistics:**

Vowels: 30-400ms (mean: 95ms)

Consonants: 20-250ms (mean: 68ms)

Words: 150-800ms (mean: 335ms)

**Pitch Statistics:**

F0 mean: 118 Hz

F0 range: 75-180 Hz

Standard deviation: 22 Hz

# Conclusion

The project delivered three major components: automated processing scripts totaling over 1,200 lines of documented Python code, a comprehensive quality assessment identifying that most flagged anomalies represent natural speech phenomena rather than errors, and systematic acoustic model comparison revealing convergence on identical alignments for high-quality data. The findings demonstrate that forced alignment technology has reached production readiness

for phonetic research applications. The 95%+ reduction in processing time compared to manual annotation enables researchers to analyze substantially larger corpora, while the high accuracy rate (fewer than 0.2% true errors) means automated outputs can be used directly for most analyses. The successful training of a custom G2P model proves that the system can be extended to specialized vocabulary domains beyond general English. Quality assurance emerged as a critical component of automated workflows. While the overall error rate was low, the 62 flagged potential issues (6.07% of phonemes) required interpretation to distinguish true errors from natural phonetic phenomena. The automated quality checker developed here successfully categorized issues by type and severity, enabling efficient manual review focused on genuinely problematic segments rather than systematic verification of all alignments. The acoustic model comparison yielded an unexpected but pedagogically valuable finding: two different models produced identical alignments on this corpus. This result suggests either that the models represent equivalent training or that high-quality recordings allow multiple models to converge on the same optimal solution. While not supporting the initial hypothesis that models would differ, this finding provides important guidance for researchers: model selection may matter less than expected for clean, clearly articulated speech. The automation scripts represent a significant deliverable with utility beyond this specific project. The modular design allows researchers to use individual components (validation only, measurements only) or execute the complete pipeline with a single command. The comprehensive documentation and clear code structure make the tools immediately usable by other researchers and valuable as educational resources for learning MFA integration with Python.

https://montreal-forced-aligner.readthedocs.io

https://kaldi-asr.org

https://www.fon.hum.uva.nl/praat/