Towards an Automatic Evaluation Model for Glottal Stop Productions: A Case Study of the Cou Language

Anonymous ACL submission

Abstract

Motivated by future needs for Indigenous language education and preservation in Taiwan, this study proposes a reproducible, end-to-end automated workflow to assess the production quality of the glottal stop (/ʔ/) in Cou, a low-resource Austronesian language of Taiwan. The pipeline comprises automatic alignment and word-level segmentation with a dockerized Montreal Forced Aligner, expert annotation, batch acoustic feature extraction via Praat/Parselmouth, and Random Forest–based binary classification with statistical testing. Empirically, the distributions of (i) F0 variation and CPP and (ii) the intensity-based deltaInt and the spectral-tilt measure H1–H2 substantially overlapped, yielding no significant differences; accordingly, the model's performance on the test set was only modest. Despite this modest performance, the workflow offers a practical example and a comparable baseline for subsequent acoustic analyses of low-resource languages and for future extensions toward event-level and glottal-source measures.

1 Introduction

The Cou people are one of Taiwan's indigenous groups, and their language belongs to the Austronesian family with a Latin-based script. Despite its typological importance, Cou has been heavily influenced by Mandarin and now faces severe decline, with fewer than 2,000 fluent speakers, rendering it critically endangered. Cou is classified within the Tsouic subgroup alongside Kanakanabu and Saaroa, though they are mutually unintelligible. It is further divided into three dialects: Tapangu, Tfuya, and Duhtu (Lee, 1979). Typologically, Cou exhibits verb-initial VOS word order and a rich focus system, contrasting sharply with the SVO order of English and the SOV order of Japanese and Korean. Phonologically, it also possesses rare implosive consonants. Earlier studies include Lee's phonological comparison (1979), Tung's descriptive grammar (1964), and Zeitoun's *Tsou Grammar* (2000).

As a critically endangered language, Cou represents an urgent case for preservation and revitalization. To this end, this study turns to computational methods to support documentation and analysis.

2 Literature Review

2.1 The Glottal Stop in Cou

In Cou, the symbol "'" corresponds to the IPA glottal stop /ʔ/, whereas h denotes the glottal fricative, indicating that the language possesses both a stop and a fricative phoneme at the glottal place of articulation (Zeitoun, 2000). The basic syllable structure is C(C)V, with a maximal template of C(C)V(V)C(C)V(V); /ʔ/ frequently occurs word-initially (e.g., 'a, 'ua) and at morpheme boundaries in compounds (e.g., 'ua-chumu). In continuous speech, unaspirated stops often exhibit short closures and low energy and can therefore be confused with silence or sonorants, making careful differentiation essential for acoustic annotation and automatic processing (Sung, 2006). Functionally, /ʔ/ is both an independent phoneme and a common pre-glottalizing element of voiced stops; in morphological derivation such as reduplication, insertion of "'" marks syllable boundaries and prevents vowel sequences. Although a stop followed by /h/ may perceptually resemble aspiration, Cou does not exhibit an aspirated–unaspirated phonemic opposition. Taken together,

/ʔ/ occupies a crucial interface between phonology and morphophonology; to evaluate its "quality" and develop automatic classification, we therefore complement a reproducible word-level baseline with event-level and glottal-source measures (e.g., GCI, open/closed-phase ratios, spectral tilt).

In current scholarship on low-resource languages, reproducible audio analysis pipelines are largely lacking, which makes procedures for data preprocessing, alignment, and feature extraction difficult to compare and hinders replication across studies. This gap yields inconsistent baselines for quality evaluation and automatic classification, thereby limiting the scalability of research across languages and speakers.

2.2 Acoustic Analysis of the Glottal Stop

Speech acoustics is often modeled by the source–filter framework (Fant, 1970; Bozkurt et al., 2007). In this framework, glottal excitation is the source and the vocal tract is the filter that shapes the measurable speech signal. Laryngeal settings shape the glottal-flow waveform and motivate parameterized studies of glottal events—e.g., the Glottal Closure Instant (GCI) and spectral tilt (Henrich et al., 2004; Alku, 2011). Traditional pipelines enhance GCI with linear-prediction residuals (LPR) or impulse-like signals, followed by peak picking and heuristic denoising; these methods work well for clean speech (Bozkurt et al., 2004). Recent CNN models that combine raw waveforms with LPR often outperform classical methods for GCI detection in pathological voices (Reddy et al., 2021). Glottal-flow estimation via inverse filtering—e.g., quasi-closed phase (QCP)—recovers the flow and yields features such as open/closed-phase ratios, closing-phase slope, and spectral tilt, benefiting pathology detection (Drugman & Alku, 2011; Airaksinen et al., 2014). These features—GCI, closing-phase slope, F0 cycle length, and open/closed-phase ratios—are quantifiable indicators of glottal-stop production (Henrich et al., 2004). In this study we consider GCI, closing-phase slope, F0-related period measures, and time-/frequency-domain parameters from glottal inverse filtering (e.g., open-/closed-phase ratios, spectral tilt). These serve as intuitive quantitative candidates for comparing glottal-source behavior before and after the realization of a glottal stop.

LPR-based GCI detection is sensitive to noise and voicing transitions, and inverse-filtering assumes a relatively stable vocal-tract filter—conditions rarely satisfied in fluent, low-resource speech. Moreover, vowel identity and recording conditions can mask these measures, yielding overlapping distributions. This helps explain our modest baseline (0.58 accuracy, 0.57 ROC-AUC); accordingly, we treat the five traditional measures as a reproducible baseline and motivate event-level, glottal-source features (e.g., GCI reliability, closing slope, open/closed-phase ratios, multi-band spectral tilt) in subsequent work.

2.3 Speech Synthesis and Applications in Low-Resource Languages

In low-resource languages, cross-lingual transfer for forced alignment can efficiently establish phoneme boundaries. For Upper Sorbian, a study applied a German acoustic model for forced alignment and treated word-initial vowel–preceding glottal stops as optional; a generalized linear mixed model (GLMM) then tested speaker-type and vowel-quality effects on glottal-stop occurrence (Pompino-Marschall & Żygis, 2010). In Choctaw, a low-resource North American language, a Kaldi-based ASR pipeline yielded relatively high WER yet demonstrated the feasibility of a "data-pipeline + base-model" framework for revitalization-oriented studies (Gordon et al., 2020). Additionally, another team used word2vec and Linguistica for corpus mining to expand lexical resources, further supporting language-technology development in low-resource settings.

However, cross-lingual alignment and synthetic-speech usage introduce domain mismatch and error propagation, which can destabilize intensity- and F0-based measures and blur /ʔ/-related contrasts. Accordingly, we disclose data-cleaning and missing-value policies, stratify evaluation by natural vs. synthetic speech, and treat alignment noise as a confound in analysis. We therefore use the current word-level pipeline as a reproducible baseline, while motivating event-level and glottal-source features in subsequent extensions to reduce domain-mismatch sensitivity.

3 Research Methodology

This study aims to develop a reproducible automated pipeline that processes both natural Cou speech and AI-generated Cou speech. The workflow includes sentence-level alignment, word-level segmentation, manual binary

annotation of glottal stop quality (0 = low quality, 1 = standard), batch extraction of five acoustic features using Praat/Parselmouth, and the training of a Random Forest model to automatically evaluate glottal stop quality.

The speech corpus used in this study was collected under a government-funded initiative on indigenous-language technologies in Taiwan. Program and project names are withheld to maintain double-blind review. We use a curated Cou speech subset from this initiative. Between 2024 and 2025, the project collected and trained models on approximately 71,000 Cou language data items. For this research, the natural speech data came from the "Cou Oral Narratives" category and featured a male speaker of the Tfuya dialect. Recordings were made using a StreamEye RODE PodMic USB dynamic microphone. The "Cou Oral Narratives" dataset contains 194 sentences, of which 154 sentences were selected for this study. For the synthetic speech corpus, Whisper large-v3-turbo was employed for automatic speech recognition (ASR), while the VITS2 model was used for text-to-speech (TTS) synthesis.

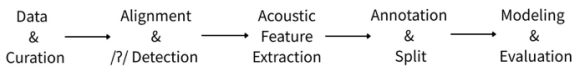

Figure 1. Research Methodology

3.1 Corpus Cleaning and Manual Annotation

The audio data were first paired and encoded with their corresponding transcriptions. To avoid compatibility issues on macOS (Apple Silicon), Docker Desktop was employed as the containerization tool. A Montreal Forced Aligner (MFA) was used for speech alignment and executed with the setting platform linux/amd64. Running MFA within Docker isolates system dependencies and prevents conflicts; using the linux/amd64 platform circumvents Kaldi/kalpy dependency errors on Apple Silicon, ensuring full reproducibility for users across operating systems. The Python environment was managed via Anaconda (Python 3.12) and JupyterLab, utilizing packages such as pandas, numpy, tqdm, soundfile, textgrid, and parselmouth for corpus segmentation and processing.

The corpus comprised two major components: natural Cou speech recordings and AI-generated speech, designed to capture both natural pronunciation and controlled acoustic conditions, thereby enhancing sample diversity and representativeness. All speech materials targeted the Cou language and were pre-segmented and quality-checked to ensure stable acoustic feature extraction. Manual annotation was conducted by trained phoneticians. Annotation guidelines specified three categories for glottal stops:

0 – Standard glottal stop: Clear glottal closure with good acoustic quality.

1 – Low-quality glottal stop: Includes incomplete closure, breathy interference, or recording distortion affecting acoustic integrity.

2 – Exclude: Samples with significant noise, segmentation errors, or indeterminate glottal stop quality.

Because a single Cou word may contain multiple glottal stops (e.g., e'a'ausnasi, o'alas'a), each occurrence was annotated separately. For example, if the first stop was standard and the second low-quality, it was annotated as "0,1," and so forth. Words containing three or more glottal stops were generally classified as excluded, and such composite annotations were reported in descriptive statistics but omitted from acoustic feature labeling.

Source	0	1	2	Total
human	417	87	90	594
synthetic	86	458	78	622

Table 1. Descriptive Statistics of Original Annotations

In the data cleaning and feature selection stage, only samples with a single annotation of either 0 or 1 were retained to ensure clear classification. The acoustic features used in glottal stop classification included closure duration, percentage change in fundamental frequency (F0) before and after closure, intensity difference across closure (deltaInt), variation in H1–H2 across closure (H1H2), and pre-closure Cepstral Peak Prominence (CPP)—a total of five features. To prevent statistical bias caused by missing values, each sample was required to contain at least three non-missing features out of the five. Any remaining missing values were either imputed using the mean or excluded in specific analyses, depending on the analytical context.

To develop an automated system for assessing glottal-stop quality, we first applied mean imputation to address missing values, performed a

stratified train–test split based on the final label (label_clean), and standardized continuous features via Z-score normalization. Five core acoustic variables—closure duration (closure), percent change in fundamental frequency (F0), cross-closure intensity difference (deltaInt), change in H1–H2 amplitude (H1H2_dB), and pre-closure cepstral peak prominence (CPP)—were summarized with groupwise descriptive statistics to characterize their distributions. Inferential testing followed a two-stage decision procedure: normality was evaluated with the Shapiro–Wilk test; if both groups approximated normality, we used independent-samples t-tests to compare means, otherwise Mann–Whitney U tests to compare distributional location.

3.2 Feature Extraction and Measurement

Acoustic feature extraction was performed using Python 3.12 and Praat, with automation through the Parselmouth interface to ensure analytical consistency and reproducibility. All audio files were pre-segmented and manually annotated in the previous stage, and analyzed at a 16 kHz sampling rate. For each token containing a glottal stop, the following five core acoustic features were extracted:

Feature	Description
closure	Duration of the glottal closure (milliseconds), approximated by the segment of minimum intensity in the intensity contour.
F0	Percentage change in fundamental frequency before and after closure, calculated as the relative difference between pre- and post-closure mean F0.
deltaInt	Intensity difference across closure, representing the change in glottal energy by comparing pre- and post-closure mean intensity.
H1H2	Variation in the H1–H2 amplitude difference across closure, reflecting glottal waveform tilt and resonance energy difference.
CPP	Cepstral Peak Prominence immediately before closure, evaluating the periodicity of glottal excitation and overall voice quality.

Table 2. Description of the Five Acoustic Features

The automated processing pipeline then proceeded as follows:

Audio loading and quality check: Automatically read each annotated and segmented .wav file and calculated the total duration of each segment.

Short-segment exclusion: Any file shorter than 0.13 seconds was skipped and logged, since Praat cannot perform reliable intensity analysis on such brief segments.

Closure detection: Identified the minimum point of the intensity contour as an approximation of the glottal closure boundaries, then set analysis windows before and after the closure to compute acoustic features.

Feature computation: Fundamental frequency (F0) and intensity were measured using Praat's to_pitch and to_intensity functions, while H1–H2 and CPP were derived through spectral and cepstral analysis.

After preliminary annotation, a total of 1,071 valid samples were obtained, including 497 tokens labeled as 0 (standard glottal stop) and 543 tokens labeled as 1 (low-quality glottal stop)—numbers that differ from the original annotation statistics. A small number of tokens contained multiple or exclusionary labels, such as those marked "2" or composite annotations (e.g., "1,0" or "0,1"). While these were reported descriptively, they were excluded from subsequent statistical analyses and machine-learning training to ensure consistency and validity.

In later stages, a condition verifying the success of feature extraction was added. Consequently, the actual number of samples entering the final analysis differed from the initial annotation totals. To maximize usable data, the automated process skipped files that were too short or failed in Praat computation or feature extraction. Most of the discarded files involved words with two glottal stops where one was standard and the other low-quality.

During statistical processing, a "minimum three-feature" condition was applied. However, because the percentage change in fundamental frequency across closure (deltaF0pct) and pre-closure Cepstral Peak Prominence (CPP_pre_dB) exhibited high rates of missing data, a mean-imputation strategy was adopted to preserve as many samples as possible. These factors resulted in final sample counts that differed from the original annotation totals.

3.3 Random Forest Classification

This study employed a Random Forest classifier implemented in the scikit-learn library as a binary

model to distinguish standard glottal stops from low-quality glottal stops. To ensure the model could handle missing data and avoid the influence of differing feature scales, data preprocessing included imputing missing values with the mean of each corresponding feature column and standardizing all features using StandardScaler to achieve a distribution with mean 0 and standard deviation 1. The dataset was stratified by label and randomly split into a 75 % training set and a 25 % test set, with 803 training samples and 268 test samples. The Random Forest model was configured with 300 trees, no depth limitation, and class_weight="balanced" to address class imbalance and enhance generalization. To further assess robustness, a stratified 5-fold cross-validation was performed across the complete dataset to obtain the mean accuracy and verify model stability. Model performance was evaluated using multiple metrics: accuracy to represent overall classification correctness, precision and recall to assess the predictive precision and sensitivity for low-quality glottal stops (class 1), F1-score as the harmonic mean of precision and recall, and ROC-AUC to measure the overall ability of the model to discriminate between the two classes.

For classification, we trained a Random Forest model (n_estimators = 300; class_weight = "balanced") and examined feature_importances_ to quantify variable contributions. Model performance on the held-out test set was evaluated using accuracy, precision, recall, F1 score, and the confusion matrix, with ROC–AUC additionally reported; five-fold cross-validation provided mean accuracy to assess generalization.

4. Results

4.1.1 Statistical Analysis

After data cleaning, which retained only samples with a single 0/1 label and at least three non-missing features, a total of 1,071 valid samples were obtained, comprising 524 tokens labeled as 0 (standard glottal stop) and 547 tokens labeled as 1 (low-quality glottal stop).

Feature	Valid count 0	Valid count 1	missing 0	missing 1
closure	524	547	0	0
F0	232	243	292	304
deltaInt	0	0	524	547
H1H2	524	547	0	0
CPP	302	313	222	234

Table 3. Feature-Wise Data Availability and Missing Counts

For the closure feature, all 1,071 tokens were fully preserved but exhibited no variation (30 ms, SD = 0), rendering statistical testing impossible. F0 and CPP showed high missing rates and no significant group differences, while deltaInt contained no valid data. H1–H2 was fully available but also demonstrated negligible effects.

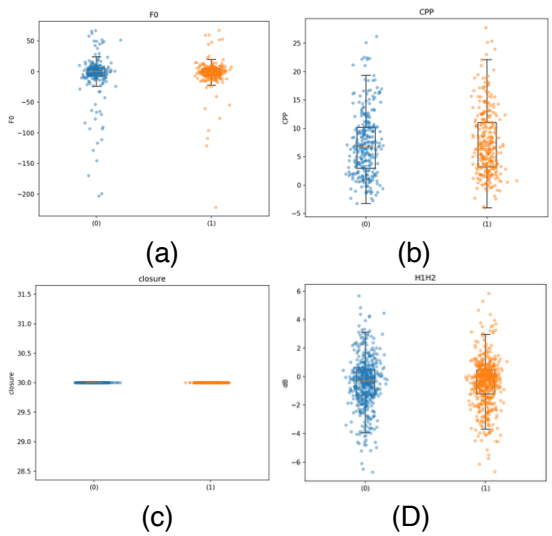

Figure 2. Scatter and box plots for CPP, F0, closure, and H1–H2. The two labels show substantial overlap; closure is near-constant (~30 ms).

4.1.2 Inter-feature Correlations

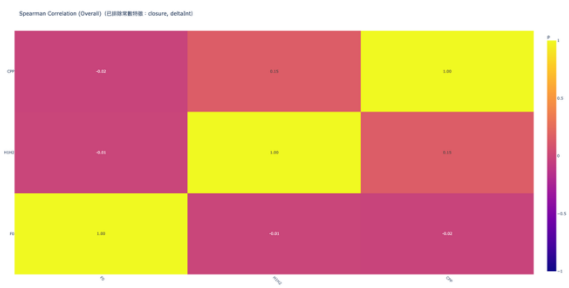

Figure 3. Spearman correlation heatmap (overall) Features with insufficient variance (closure, deltaInt) were excluded; aside from a weak H1H2–CPP association (ρ≈0.15), most pairwise correlations are near zero, indicating low inter-feature redundancy.

To assess feature coupling, we computed the overall association matrix using Spearman's rank correlation coefficient (ρ), a robust measure under non-normality and outliers. The results indicate generally weak inter-feature associations: H1H2–

CPP shows a weak positive correlation (approximately ρ = 0.15), likely reflecting their shared sensitivity to source-quality properties (e.g., spectral tilt and regularity of vocal-fold vibration). Correlations between F0 and the other features are near zero (|ρ| < 0.05), suggesting limited linkage between prosodic variation and source-quality indices. This pattern implies low redundancy among the features.

4.2 Random Forest Performance Evaluation

A binary classification experiment was conducted using the cleaned dataset, with the additional requirement that all five acoustic features be non-missing. For the feature with a high missing rate, deltaInt, missing values were imputed with the mean. The final dataset comprised 676 usable samples, including 335 tokens labeled 0 (standard glottal stop) and 341 tokens labeled 1 (low-quality glottal stop), resulting in nearly balanced classes. The data were stratified by label and split into a 75 % training set (507 samples) and a 25 % test set (169 samples). To minimize the impact of missing values, feature-level mean imputation was applied, followed by StandardScaler normalization to a mean of 0 and a standard deviation of 1. Features were mean-imputed and Z-score standardized prior to training.

Model performance on the test set was evaluated using accuracy, precision, recall, and F1-score, and further assessed with Receiver Operating Characteristic (ROC) curves and Precision–Recall (PR) curves, reporting the area under the ROC curve (ROC-AUC) and Average Precision (AP). A confusion matrix was also presented to examine the types of misclassifications.

Metric	value
Accuracy	0.580
Precision	0.581
Recall	0.588
F1-score	0.585
ROC-AUC	0.572
Average Precision (PR)	0.56

Table 5. Random Forest Model Performance Metrics

(a) Overall and Class-Level Metrics

Taking the low-quality glottal stop as the positive class, performance was modest across all evaluation metrics: Precision = 0.581, Recall = 0.588, and F1-score = 0.585. In addition, the ROC-AUC value of 0.572 and the Average Precision (AP, area under the Precision–Recall curve) of 0.56 both indicate that the classifier performed only slightly better than chance. These results suggest that the current acoustic features provide limited discriminative power for distinguishing standard from low-quality glottal stops.

(b) Confusion Matrix

Specifically, among the 84 samples truly labeled as standard, 48 were correctly predicted as standard, while 36 were misclassified as low-quality. Conversely, among the 85 samples truly labeled as low-quality, 50 were correctly predicted as low-quality, but 35 were misclassified as standard.

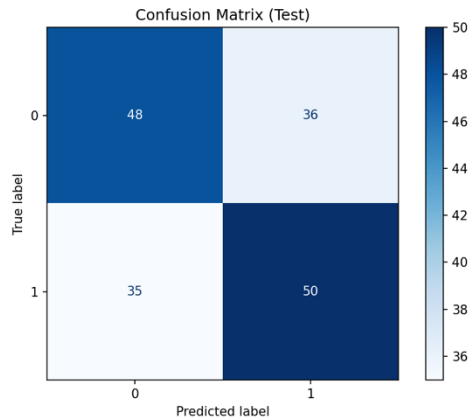

Figure 4. Confusion Matrix (Test Set)

The matrix shows broadly symmetric performance across classes. misclassifications are comparable between classes, indicating no pronounced class-specific bias.

(c) Curve Analysis

The ROC curve (Figure 5) shows an AUC of 0.572, only slightly higher than the random-classification baseline of 0.5. The Precision–Recall (PR) curve (Figure 6) has an Average Precision (AP) of 0.56; in the high-recall region (recall ≳ 0.6), the precision remains around 0.5–0.6, likewise reflecting substantial overlap between the two class distributions.

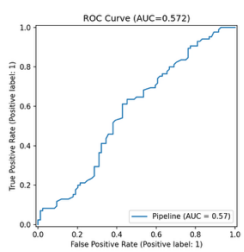

Figure 5

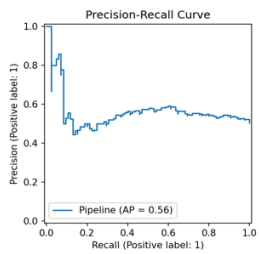

Figure 6

Figure 5. ROC Curve of the Random Forest Model (AUC = 0.572)

Figure 6. Precision–Recall Curve of the Random Forest Model (Average Precision = 0.56)

(d) Overall performance—accuracy of about 0.58 and ROC-AUC of about 0.57—demonstrates that binary classification based on the five extracted acoustic features provides only limited discriminative power.

Because closure is a constant feature, its theoretical importance approaches zero; the model relied primarily on features such as H1H2 and CPP, which provided only minimal information gain—insufficient to achieve strong separability. Possible contributing factors include the constant nature of closure (offering no information), the high missing-rate and instability of some features, and the likelihood that current broad time–frequency indicators fail to capture subtle differences in data quality.

Future improvements should include: (1) finer-grained annotation, providing precise onset and offset times for each /ʔ/ and using a single glottal-stop event as the unit of analysis; (2) feature expansion, incorporating measures such as glottal flow estimation, spectral tilt, and noise ratios, or applying speaker-internal normalization; and (3) enhanced evaluation design, such as speaker-independent cross-validation (e.g., GroupKFold) and comparison with alternative ensemble methods like XGBoost or LightGBM to explore potential upper limits of model performance.

4.3 Acoustic Interpretation

The limited discriminability observed in Section 4.2 is consistent with properties of the current feature set and annotation design. Closure is quasi-constant (~30 ms), reflecting the fixed analysis window rather than physiological variability. F0 is strongly shaped by prosodic context (intonation, stress, speech rate), reducing its utility for quality judgments. DeltaInt exhibits extensive missingness and unstable SNR. H1–H2 and CPP provide only small effects, and in natural speech their cues can be masked by vowel resonances, further weakening source-quality contrasts. Finally, because labels were assigned at the word level, tokens containing multiple /ʔ/ events were collapsed to a single binary label, diluting event-level variation and obscuring transient acoustic differences.

5. Conclusion

The primary contribution of this study is the construction of a reproducible, end-to-end automated pipeline for assessing the production quality of the glottal stop (/ʔ/) in Cou, together with the systematic integration of natural and synthetic speech corpora. The pipeline encompasses dockerized automatic alignment and word-level segmentation, an auditable expert-annotation workflow, batch acoustic feature extraction via Praat/Parselmouth, and machine-learning evaluation coupled with statistical testing. Methodologically, we provide (1) a comparable word-level baseline and transparent data-processing specifications that facilitate cross-study replication and side-by-side comparison; (2) empirical evidence for the limitations of the five traditional features on this task, thereby motivating the necessity of event-level and glottal-source measures; and (3) a reproducibility practice for low-resource speech research through principles of natural/synthetic stratification, missing-data handling, and open configuration. Overall, beyond reporting experimental observations, this paper establishes a methodological benchmark and presents an open workflow that offers a reusable starting point and comparative baseline for acoustic research and language-technology development in low-resource Austronesian languages, while providing a clear technical pathway for the subsequent incorporation of glottal-source measures such as GCI, phase ratios, and spectral tilt.

6. Limitations

Future work will prioritize event-level annotation of /ʔ/ with precise onset/offset boundaries, speaker-independent data partitioning, and a balanced composition of natural and synthetic materials, accompanied by explicit inter-annotator agreement metrics and systematic error tracking. Beyond F0, intensity, and H1–H2, we will broaden the feature space to glottal-source and signal-quality measures (e.g., GCI, open/closed-phase ratios, spectral tilt, HNR, glottal flow, EGG), together with speaker-internal normalization and compensation for recording conditions to enhance stability and comparability. Modeling will examine Gradient Boosting, XGBoost, and sequence-based neural networks, integrating uncertainty estimation and error-bucket analyses, and validating with cross-corpus and cross-lingual protocols to strengthen external validity. Collectively, this roadmap couples event-level labeling with glottal-source

features and rigorous reproducibility standards to improve effectiveness, generalizability, and the cumulative comparability of research on low-resource Austronesian languages.

7. References

Boersma, P., & Weenink, D. 2023. Praat: Doing phonetics by computer (Version 6.4.14) [Computer software]. University of Amsterdam.

Jadoul, Y., Thompson, B., & de Boer, B. 2018. Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics, 71:1–15.

Brixey, J., & Traum, D. 2022. Towards an automatic speech recognizer for the Choctaw language. In 1st Workshop on Speech for Social Good (S4SG), Incheon, Korea, September 24–25.

Brixey, J., Sides, D., Vizthum, T., Traum, D., & Iskarous, K. 2020. Exploring a Choctaw language corpus with word vectors and minimum distance length. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 2746–2753. European Language Resources Association.

Drugman, T., Alku, P., Alwan, A., & Yegnanarayana, B. 2020. Glottal source processing: From analysis to applications. Computer Speech & Language. Preprint. https://arxiv.org/abs/1912.12604

Kraljevski, I., Bissiri, M. P., Duckhorn, F., Tschoepe, C., & Wolff, M. 2021. Glottal stops in Upper Sorbian: A data-driven approach. In Proceedings of Interspeech 2021, pp. 1101–1105.

M, G. R., Mandal, T., & Rao, K. S. 2018. Glottal closure instants detection from pathological acoustic speech signal using deep learning. In Machine Learning for Health (ML4H) Workshop at NeurIPS 2018.

Narendra, N. P., & Alku, P. 2020. Glottal source information for pathological voice detection. IEEE Access, 8:1–13.

Perrotin, O., & McLoughlin, I. V. 2020. Glottal flow synthesis for whisper-to-speech conversion. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:889–901.

Chun-chih, L. 2014. A study on the sound changes of stops in Chinese historical phonology. Journal of Chinese Literature, NCCU, 21:163–212.

Chien, C.-Y. 2007. A syllable duration model for Mandarin speech (Master's thesis). National Chiao Tung University.

Liu, K.-Y. 2011. An implementation of HMM-based English speech synthesis (Master's thesis). National Chiao Tung University.

Zhou, W.-S., & Huang, Y. 2019. Acoustic analysis of Korean students' acquisition of Chinese stop consonants. Journal of Chinese Language and Literature, 128:145–162.

Wei, L.-Y. 2011. The production of stop consonant in Mandarin children with a cochlear implant: Analysis of voice onset time and accuracy of stop production (Master's thesis). University of Taipei.