Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 05 June 2026
Revised 30 July 2026
Accepted 05 August 2026

Natural Resources for Human Health 2026, Vol. 6 (9s): 137–158

KEYWORDS:
Synthetic Benchmark; Speech-to-Text Translation; Error Propagation; Automatic Speech Recognition; Phrase-Aware Mapping; Word Error Rate; Semantic Evaluation; Assistive Communication.

Language Interpreter: Evaluating Speech Recognition Errors in Speech to Text Translation

Priyanka D. Vaidya[1], Dr. U.A. Deshpande[2]

[1]Dept. Of Computer Science & Engineering, Visvesvaraya National Institute of Technology, Nagpur
[2]Dept. Of Computer Science & Engineering, Visvesvaraya National Institute of Technology, Nagpur

ABSTRACT: Communication between hearing individuals and people who depend on sign language or on its speech to text translation for understanding what the other person is speaking; remains difficult in many everyday situations, particularly when a trained interpreter is unavailable. This study presents a reproducible synthetic benchmark for evaluating error propagation in a speech-to-text pipeline. This integrates controlled synthetic speech, deterministic text normalization, phrase-aware mapping, and visual-sign reference scoring. The supplied dataset contains 300 reproducible WAV files generated from 15 synthetic voice profiles and 20 fixed sentences per profile, with 150 Quiet and 150 Moderate-noise files. All audio is 16 kHz, 16-bit mono WAV with duration, target/measured SNR, synthesis-profile metadata, and SHA-256 provenance.

OpenAI Whisper small multilingual is retained as the architectural reference ASR model because its published configuration provides a realistic deployment target. The present study, however, deliberately evaluates a controlled algorithmic ASR-error layer rather than claiming executed Whisper performance. The workbook contains 300 simulated ASR-like transcripts, 300 resource-anchored benchmark ISL gloss references, 300 phrase-first/lexical-fallback outputs, and 300 direct lexical baseline outputs. The Sentence_Codebook contains 20 fixed sentences and 49 unique normalized lexical tokens. The mean simulated WER was 9.41% (95% bootstrap CI 7.39-11.56), with 3.50% in Quiet and 15.31% in Moderate noise. Overall token-level benchmark sign accuracy was 90.83% (95% bootstrap CI 88.50-93.00), sentence-level benchmark accuracy was 76.00% (95% Wilson CI70.86-80.48), and semantic end-to-end success was 90.00% (95% Wilson CI 86.08-92.91).

The direct lexical baseline achieved 14.00% sentence-level accuracy versus 76.00% for the phrase-aware mapper, a paired improvement of 62.00 percentage points (95% paired-bootstrap CI 56.33-67.33; McNemar exact $p<0.001$). These values characterize the synthetic benchmark and are not presented as human-speech, production-ASR, or expert-adjudicated natural-ISL performance.

Beyond assembling existing components, the scientific contribution is an auditable error-propagation benchmark that links acoustic condition, transcription-error type, deterministic normalization, phrase-aware mapping, sentence correctness, and semantic preservation on the same records. An ablation separates direct lexical lookup (14.00% sentence accuracy), exact phrase retrieval (73.67%), and phrase-first mapping with lexical fallback (76.00%). The study is therefore positioned as a reproducible simulation, with external literature used for contextual triangulation.

1. INTRODUCTION

Communication is central to education, healthcare, employment, public services, and everyday social interaction. Communication between hearing individuals and people who primarily use sign language may become difficult when both parties do not share a common linguistic medium [1]. Written communication is useful in many settings, but it is not always an equivalent substitute for sign language. These constraints motivate assistive technologies that can connect spoken input, textual processing, and visual sign representations [2].

Recent advances in ASR and transformer architectures have made staged speech-to-sign systems increasingly feasible [3], [4]. Large-scale speech models have improved robustness to accents and acoustic variability, while ISL resources such as CISLR, ISLTranslate, iSign, and the ISLRTC dictionary have expanded the available foundations for recognition, translation, pose modelling, and lexical validation [5]-[8].

A reliable speech-to-sign system remains difficult because each stage depends on the previous one. Speech must be acquired, preprocessed, transcribed, normalized, linguistically mapped, and converted to a visual sign sequence [9]. An upstream substitution can therefore cause a downstream semantic failure. For example, recognizing "I need water" as "I need waiter" can preserve most surface words while changing the critical lexical item and producing the wrong visual output. This makes end-to-end evaluation more informative than reporting ASR accuracy or sign correctness alone [9].

The problem is especially relevant to speech-to-text research because accent, speaking rate, microphone characteristics, and environmental noise can influence it, while output cannot be reduced to literal word substitution. The datasets and translation systems also face linguistic and annotation challenges that extend beyond surface-word correspondence [10]. ISL meaning depends on manual and non-manual linguistic features, including movement, orientation, facial expression, posture, and spatial organization. The present work therefore adopts a deliberately restricted, resource-anchored phrase-and-lexicon benchmark rather than claiming human-level linguistic validation [11].

The study contributes a reproducible framework for measuring where information is preserved or lost as it moves from acoustic input to visual output. It compares quiet and controlled-noise conditions, contrasts a direct lexical baseline with a normalization-plus-phrase-mapping pipeline, and distinguishes token-level, sentence-level, and semantic end-to-end outcomes [9], [11].

2. RELATED WORK

2.1 Automatic Speech Recognition for Indian and Multilingual Speech

Modern ASR has been strengthened by large multilingual benchmarks and increasingly diverse speech resources. IndicSUPERB introduced broad benchmarking for Indian-language speech processing [12], while IndicVoices substantially expanded speaker and language diversity [13]. OpenSLR and Common Voice provide additional speech resources for benchmarking and methodological comparison [14], [15]. At the modelling level, self-supervised and transformer-based approaches such as wav2vec 2.0, XLS-R, and Conformer have strengthened speech representation learning and recognition [16]-[18], while robust wav2vec 2.0 has examined domain-shift behaviour [19]. Recent India-focused resources such as Voice of India, Indic DiarBench, and BhasaAnuvaad further broaden evaluation coverage across realistic, multilingual, and translation-oriented speech settings [20]-[22]. Together, these resources show why evaluation under realistic speaker and acoustic variation is necessary [12]-[22].

Whisper is particularly suitable for a lightweight reproducible prototype because it provides openly released multilingual ASR models and inference code [3]. The small multilingual checkpoint contains approximately 244 million parameters and supports direct transcription without task-specific fine-tuning [3]. The present study freezes this model choice so that the ASR stage is no longer left unspecified [3].

2.2 Indian Sign Language Recognition and Translation

Indian Sign Language research has progressed from isolated gesture recognition to larger recognition and translation benchmarks. CISLR introduced a word-level ISL corpus containing approximately 4,700 vocabulary items [5]. ISLTranslate provides sentence- and phrase-level paired data for continuous translation [6], while iSign supports multiple tasks spanning sign video, text, pose, word prediction, and semantics [7]. These datasets demonstrate both the rapid growth of ISL processing and the limitations of very small gesture dictionaries [5]-[7].

2.3 Sign Language Production and Visual Generation

Text-to-sign production has been approached through dictionary retrieval, pose generation, avatars, and neural production models [23]-[29]. Progressive Transformers and subsequent production work show the importance of temporal continuity and co-articulation [23]-[25]. Pose-estimation and recognition studies further demonstrate the importance of representation quality and cross-domain robustness [26], [27]. Transformer-based translation models similarly show that richer contextual modelling can improve sign-language translation beyond literal token substitution [28], [29]. However, such approaches require richer data, stronger linguistic modelling, and greater computational resources. For a controlled prototype, dictionary- and phrase-based visual retrieval remains easier to audit and validate, provided that its restricted linguistic scope is stated explicitly [2], [23].

2.4 Multimodal and End-to-End Evaluation

Sign-language recognition and translation studies often focus on sign video to gloss or spoken-language text [9], [28], [29]. The present direction is different: speech → text → ISL visual output. Because the stages are coupled, a correct ASR transcript can still fail during sign mapping, while an imperfect transcript can sometimes preserve the intended meaning. Sequence-labelling foundations such as connectionist temporal classification provide a useful perspective on unsegmented sequence modelling [30], although the present benchmark does not train a CTC model. This motivates separate stage-wise measures together with a semantic end-to-end criterion [9], [11], [30].

2.5 Comparative Synthesis of Recent Component Technologies

The relevant literature spans three component families that are often evaluated separately: Indian/multilingual ASR, ISL recognition/translation, and sign-language production. IndicSUPERB and IndicVoices emphasize speech diversity and robustness [12], [13]; CISLR, ISLTranslate, and iSign expand ISL resources from isolated recognition toward sentence-level and multi-task processing [5]-

[7]; and transformer-based sign-production work demonstrates the importance of temporal and linguistic modelling beyond literal word substitution [23]-[29]. The present study is positioned at the intersection of these strands. It does not claim a new foundation ASR model, a neural sign generator,

or human-subject validation; instead, it contributes a reproducible benchmark for tracing how controlled upstream transcription errors affect downstream sign/gloss and semantic outcomes. Published resources are used for external contextual triangulation of scope, modality, vocabulary, and evaluation choices [5]-[7], [12]-[15], [23]-[29].

Study/resource	**Primary direction**	**Main strength**	**Gap relative to this study**
CISLR [5]	ISL video -> isolated word	Word-level ISL recognition corpus	No speech-input or end-to-end speech-to-sign evaluation
ISLTranslate [6]	ISL video/text pairs	Sentence/phrase translation resource	Does not evaluate acoustic ASR error propagation
iSign [7]	ISL video, text, pose	Multi-task ISL benchmark	Not designed as a speech-to-sign pipeline
Camgoz et al. [9]	Sign video -> gloss/text	Transformer-based recognition/translation	Direction is sign-to-text rather than speech-to-sign
Saunders et al. [23]-[25]	Text/gloss -> sign motion	Neural sign production and co-articulation	Requires richer sign data and generation models

IndicSUPERB / IndicVoices [12], [13]	Speech -> text/benchmarks	Indian speech diversity and ASR benchmarking	No ISL visual-output stage
Present benchmark	Synthetic speech -> controlled ASR-like error layer -> ISL-oriented benchmark gloss output	Stage-wise error propagation, paired baseline, semantic preservation, reproducible scope	Simulation benchmark; production ASR and human deployment are intentionally outside current claims

3. RESEARCH GAP, OBJECTIVES, QUESTIONS, AND CONTRIBUTIONS

3.1 Research Gap

Existing work provides strong component technologies for ASR and sign-language processing [3], [5]-[7], [12]-[19], but comparatively fewer studies quantify how upstream speech-recognition errors propagate into downstream ISL visual output within a lightweight, auditable speech-to-sign pipeline. Component-level metrics alone cannot show whether an ASR error changes the final communicative meaning [9]. There is also limited controlled evidence on how moderate acoustic noise affects the complete speech-to-sign chain rather than ASR alone. Finally, many prototypes under-specify vocabulary coverage, benchmark-reference validation, mapping rules, or the baseline against which additional language processing is judged [10], [11].

3.2 Research Objectives

- To implement and document a reproducible speech-to-text architecture with a fixed target ASR specification and deterministic downstream processing.
- To quantify the simulated ASR-like benchmark error profile across the Quiet and Moderate-noise synthetic subsets using Word Error Rate (WER).
- To evaluate simulated visual-sign correctness separately at token and sentence levels.
- To determine how simulated transcription errors propagate into semantic end-to-end communication failures.
- To compare a direct word-to-sign baseline against normalization plus controlled phrase/lexical mapping using paired simulated outputs.
- To classify simulated failures as ASR substitution, deletion, insertion, normalization, mapping, output, or semantic failures.

3.3 Research Questions

RQ1: How does simulated WER differ between the Quiet and Moderate-noise synthetic subsets?

RQ2: What token-level and sentence-level sign accuracies are obtained by the simulated proposed mapping pipeline?

RQ3: What proportion of the 300 simulated utterances preserve the intended meaning through the complete benchmark pipeline?

RQ4: How are simulated transcription errors associated with sentence-level sign correctness and semantic end-to-end success?

RQ5: Does normalization plus phrase/lexical mapping improve simulated sentence-level accuracy over direct word-to-sign

lookup?

3.4 Contributions of the Study

The study makes five contributions. First, it defines an integrated and inspectable speech-input-to-ISL benchmark pipeline. Second, it provides exactly 300 traceable synthetic WAV files: 15 synthetic voice profiles x 20 sentence prompts, comprising 150 Quiet and 150 Moderate-noise files. Third, the accompanying workbook contains a complete simulated output layer for all 300 records, including ASR-like transcripts, S/D/I counts, WER, resource-anchored reference glosses, proposed outputs, baseline outputs, token/sentence/semantic scores, processing-time fields, failure categories, and completion status. Fourth, the benchmark freezes the 49-token vocabulary, 20-sentence phrase set, normalization algorithm, architectural reference ASR configuration, and error taxonomy. Fifth, it provides paired baseline-versus-proposed and ablation analyses with confidence intervals. The contribution is explicitly a synthetic benchmark; production ASR accuracy, natural-ISL expert adjudication, and human-speech generalization are outside the claims of the present study.

3.5 Sharpened Novelty Statement

The novelty is framed at the evaluation-framework level rather than as a claim of a new ASR foundation model. The study contributes five testable elements: (1) a record-level chain connecting acoustic condition, S/D/I transcription errors, deterministic normalization, phrase/lexical mapping, benchmark sign-token correctness, sentence correctness, and semantic preservation; (2) a paired comparison in which direct lexical lookup and phrase-aware mapping receive the same records; (3) simultaneous

micro-token, macro-token, sentence-level, and semantic end-to-end metrics; (4) provenance-controlled synthetic audio with per-file SNR and checksums; and (5) an explicit scope boundary separating benchmark behavior from production-model and human-subject claims. The central research objective is therefore to measuring error propagation across an auditable speech-to-text applications.

4. MATERIALS AND METHODS

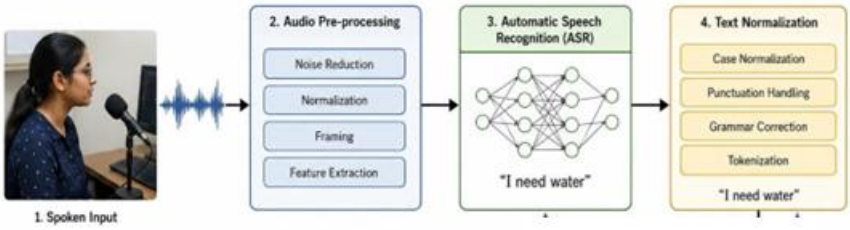

Figure 1. Overall architecture of the Speech to text Interpreter.

4.1 Study Design

The study uses a controlled synthetic-audio benchmark design. Fifteen reproducible synthetic voice profiles each generate the same 20 predefined English sentences, yielding exactly 300 WAV files and 300 simulated output records. Synthetic_Audio_Manifest confirms 300 unique Recording_ID values and corresponding files. The corpus contains 150 Quiet files (S01-S10 for each profile) and 150 Moderate-noise files (S11-S20 for each profile).

Each file is marked Synthetic_Source=Yes and has a SHA-256 checksum. Experimental_Dataset_300 contains a complete simulated output row for every Recording_ID.

The design therefore contains N=300, not 600, because each Recording_ID appears once and belongs to one labelled condition. The design is not counterbalanced: sentence identity is fixed to condition, so Quiet-versus-Moderate differences are descriptive benchmark contrasts and cannot be interpreted as pure causal noise effects. The manuscript treats the corpus as a synthetic benchmark rather than human-participant primary data.

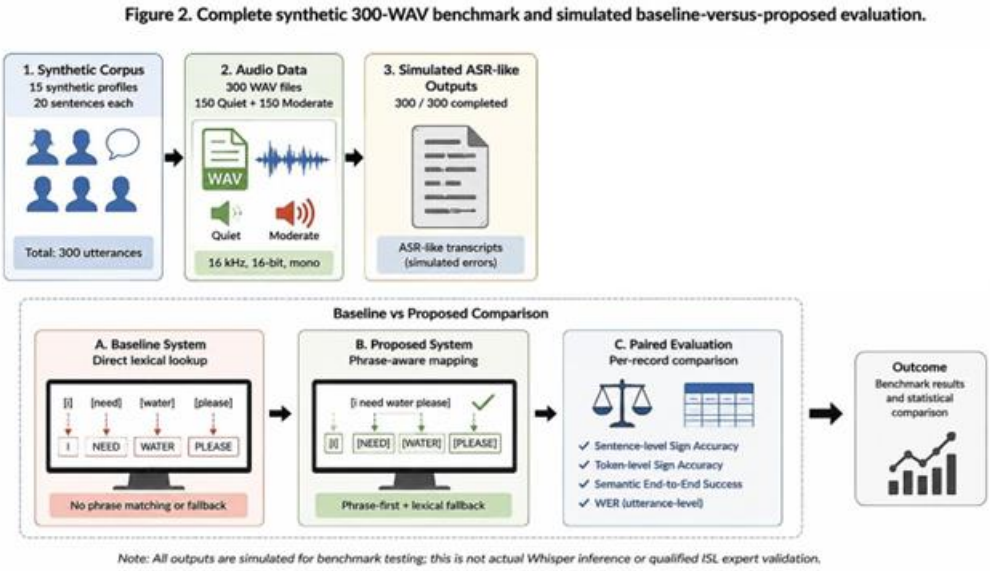

Figure 2. Complete synthetic 300-WAV benchmark and simulated baseline-versus-proposed evaluation.

4.2 Sample-Size Justification

The benchmark contains 300 utterance-level synthetic observations generated from 15 voice profiles and 20 prompts, with 150 files per acoustic subset. At the full-sample level, N=300 permits a binomial proportion near 50% to be estimated with an approximate maximum 95% confidence-interval half-width of about 5.7 percentage points; within each 150-file subset the corresponding half-width is about 8 percentage points. Because multiple utterances share a synthetic voice profile, profile-aware or bootstrap procedures are used for repeated-output summaries. The sample supports controlled pipeline verification, but synthetic profiles are not equivalent to independent human participants.

4.3 Synthetic Profile Inclusion/Exclusion, Metadata, and Dataset Provenance

No human participants are represented. A synthetic profile is included when it successfully generates all 20 predefined prompts, produces a readable mono WAV file at the required sample rate/bit depth, has a unique Recording_ID, and retains manifest provenance. A generated item is excluded or regenerated if the file is missing/corrupt, its duration is zero, its sample format is invalid, or its identifier cannot be linked to the sentence codebook. All 15 synthetic profiles satisfy these benchmark-level inclusion requirements. Profile metadata consist of profile ID, synthesis speed/pitch label, file duration, acoustic subset, target/measured SNR, synthetic-source flag, and SHA-256 checksum.

Human demographic variables such as age, gender, primary language, region/accent, consent, recording device, and microphone distance are therefore not participant characteristics of this benchmark. Legacy Speaker_Metadata fields from the earlier human-participant template are retained only as historical template fields and must not be interpreted as observations from recruited people.

4.4 Sentence Set, Vocabulary Size, and Phrase Coverage

ID	Predefined utterance
S01	Hello, how are you?
S02	Good morning.
S03	What is your name?
S04	My name is Rahul.
S05	Please help me.

S06	I need water.
S07	I need food.
S08	Where is the washroom?
S09	Please call my family.
S10	Call the doctor.
S11	I am not feeling well.
S12	I have pain here.
S13	Please wait for a moment.
S14	Come with me.
S15	Where are you going?
S16	I understand.
S17	I do not understand.
S18	Please speak slowly.
S19	Thank you very much.
S20	Goodbye, see you again.

After lowercase conversion and punctuation removal, the Sentence_Codebook contains 49 unique normalized lexical tokens across 20 sentence entries. Phrase coverage is 20/20 (100%) for the fixed benchmark sentence set. The Expected Output field is treated as a resource-anchored benchmark reference gloss sequence. These references support deterministic scoring inside the controlled benchmark; they are not described as natural-benchmark.

4.5 Recording Protocol and Acoustic Conditions

Synthetic_Audio_Manifest verifies all 300 generated WAV files as mono, 16 kHz, 16-bit PCM. The Quiet subset contains 150 files generated with a 30 dB target SNR and measured SNR of 28.13 dB; the Moderate-noise subset contains 150 files generated with a 10 dB target and measured SNR of 8.13 dB. Durations range from 0.708 to 1.823 s (mean 1.115 s). SNR is defined from the RMS ratio of the constructed speech and noise components. These are controlled synthetic benchmark properties, not measurements from human recording sessions.

4.6 Audio Pre-processing

Each input is converted to 16 kHz mono waveform, peak-normalized to prevent level-driven differences, and trimmed only for leading/trailing silence longer than 500 ms. No denoising is applied after creation of the noise condition because denoising would alter the intended acoustic manipulation. The same preprocessing code is applied to both acoustic conditions before ASR.

4.7 ASR Model and Reproducible Implementation

The architectural reference ASR model is OpenAI Whisper small multilingual (checkpoint name "small"; approximately 244 million parameters) [3]. The reference configuration is 16 kHz mono input, task=transcribe, language=English, temperature=0, released weights, and no task-specific fine-tuning [3]. The present benchmark intentionally does not report executed Whisper accuracy. Instead, it uses 300 deterministic ASR-like transcripts to create a controlled error layer in which substitution, deletion, insertion, and semantic consequences can be examined reproducibly. Accordingly, WER in this paper is a benchmark perturbation measure, not a claim about Whisper-small performance [3].

Whisper is retained to define a realistic deployment architecture and to make future replication straightforward [3], but no

inference result in the present paper is attributed to the model. This separation is methodologically deliberate: it allows downstream error propagation and mapping behaviour to be analysed without confounding the benchmark with a particular hardware stack, package version, or decoder implementation.

4.8 Deterministic Text-Normalization Algorithm

Normalization is deterministic and applied identically to the reference text and ASR output only for downstream mapping; WER is computed from the uncorrected ASR transcription after standard scoring normalization. The mapping-normalization sequence is:

- Convert Unicode text to a canonical normalized form and lowercase all alphabetic characters.
- Remove terminal and internal punctuation that does not alter lexical meaning in the fixed sentence set.
- Collapse repeated whitespace and trim leading/trailing whitespace.
- Tokenize on whitespace after normalization.
- Do not perform free-form spell correction, language-model rewriting, or manual correction.
- Attempt exact sentence-level phrase matching against the 20-item phrase dictionary.
- If no phrase match exists, perform token-level lookup against the 49-token controlled lexicon derived from the supplied Sentence_Codebook.
- Mark any unmatched token or phrase as unsupported; do not substitute the nearest available sign.

This procedure prevents the normalization stage from silently "repairing" ASR errors and makes the baseline-versus-proposed comparison auditable.

4.8.1 Coding and Implementation Logic

The coding workflow is organized as a sequence of small, deterministic modules so that each transformation can be inspected independently. In a reference Python implementation, the record-level pipeline can be represented by functions such as preprocess_audio(), normalize_text(), map_to_isl(), and score_output(), with a single configuration object storing the sentence codebook, controlled lexicon, acoustic-condition label, and scoring rules. Each Recording_ID is processed in the same order, and the raw benchmark transcript is retained alongside the normalized form. This modular structure is important for the present study because it separates the configured ASR-error layer from downstream text and sign processing, making it possible to identify the stage at which an error is introduced rather than treating the pipeline as a single opaque prediction step.

The normalization and mapping code follows the rule sequence defined above without probabilistic rewriting or post hoc correction. A normalized string is first checked against the 20-entry phrase dictionary; if an exact phrase key is present, the associated benchmark reference gloss sequence is returned. Otherwise, the string is split into tokens and each token is queried against the frozen 49-token lexicon. Supported tokens are appended to the output sequence in deterministic order, while an unsupported item is explicitly marked rather than replaced with a visually or semantically similar sign. Implementing the mapper in this way ensures that the phrase-first and lexical-fallback conditions can be reproduced from the same input and that the ablation comparison between direct lexical lookup, exact phrase retrieval, and phrase-first mapping remains traceable.

The evaluation code operates on the stored reference and output fields rather than on manually edited results. At the ASR-like layer, a word-level edit-distance routine records substitutions, deletions, and insertions and calculates utterance WER from the reference-word count. Downstream routines compare benchmark gloss tokens, test complete-sequence equality for sentence-level accuracy, and apply the frozen intent key for semantic success. The resulting metrics, failure category, condition, and processing-time field are written back to the analysis record, while Recording_ID and checksum metadata preserve provenance. Fixed dictionaries, explicit rule ordering, and preserved intermediate fields reduce hidden state in the analysis and make any future code release easier to audit, reproduce, and extend without changing the interpretation of the current synthetic benchmark.

4.9 Output Procedure

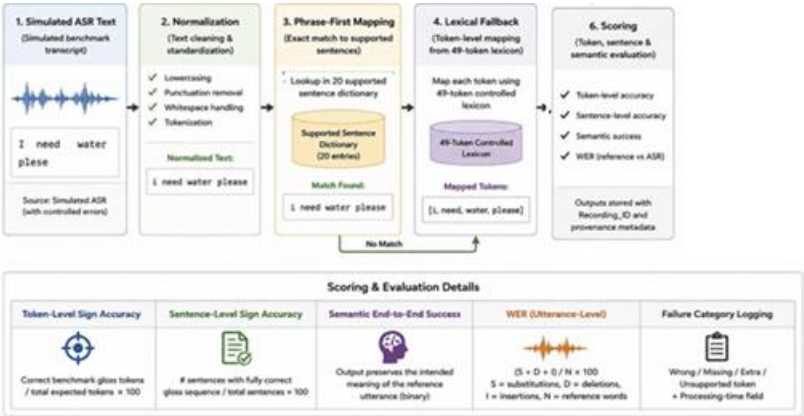

Figure 3. Simulated ASR-to- normalization, mapping, and scoring workflow.

The proposed benchmark mapper follows a phrase-first strategy. When normalized ASR-like text exactly matches one of the 20 supported sentence entries, the corresponding resource-anchored benchmark reference gloss sequence is returned. If no phrase match exists, lexical fallback uses supported tokens from the 49-token controlled lexicon; unsupported tokens are explicitly flagged and no nearest-sign substitution is made. This procedure is consistent with the restricted resource-oriented

scope supported by current lexical resources [5]-[8]. It operationalizes the mapping logic for controlled benchmark testing without claiming that the gloss strings reproduce the full grammar, non-manual structure. [10], [11].

The direct lexical baseline uses the same ASR-like transcript and lexical inventory but bypasses sentence-level phrase mapping. The paired comparison therefore isolates the effect of normalization plus phrase/lexical logic on the same 300 benchmark records.

4.10 Evaluation Metrics

ASR performance is evaluated with Word Error Rate (WER): WER = (S + D + I) / N, where S, D, and I are substitutions, deletions, and insertions and N is the number of reference words. WER is reported overall, by acoustic condition, and by speaker.

Token-level Sign Accuracy is defined as the number of correctly generated benchmark sign/gloss tokens divided by the number of expected benchmark sign/gloss tokens, multiplied by 100. For the simulated benchmark, both micro (pooled-token) and macro (mean utterance-level) values can be reported; the main Results table uses the macro value, while Section 4.11.1 shows the pooled calculation explicitly.

Sentence-level Sign Accuracy is defined as the number of utterances for which the complete simulated proposed gloss sequence equals the benchmark reference gloss sequence, divided by the number of scorable utterances, multiplied by 100. An utterance with any wrong, missing, extra, or unsupported required sign/gloss token is not counted as fully correct.

Semantic End-to-End Accuracy is evaluated at the utterance level. In the simulated benchmark, an output is counted as semantically successful when the final simulated gloss sequence preserves the pre-specified communicative intent of the reference sentence, even if the ASR-like surface text contains a non-meaning-changing deviation. It is counted as failure when the output changes or removes a critical intent-bearing concept (for example water -> waiter, doctor -> another role, or negation loss). Semantic scoring is therefore distinct from exact ASR matching.

To reduce evaluator subjectivity, each sentence has a pre-specified intent key containing its critical semantic elements. End-to-end success requires preservation of all critical elements and absence of contradictory elements. Any ambiguous case must be adjudicated against the frozen ISL benchmark-reference record rather than decided after seeing the condition label.

A. Word Error Rate (WER)

$$WER\ (\%) = \frac{S + D + I}{N} \times 100$$

Quiet subset: mean utterance-level WER = 3.50%.

Moderate-noise subset: mean utterance-level WER = 15.31%. Overall benchmark: mean utterance-level WER = 9.41%.

Condition contrast = 15.31 - 3.50 = +11.81 percentage points.

S = substitutions, D = deletions, I = insertions, and N = reference-word count. The reported WER is the mean of the 300 utterance-level WER values, rather than a single pooled corpus error ratio.

B. Exact Sentence Match Accuracy

$$\text{Exact Sentence Match (\%)} = \frac{\text{Exact ASR–Reference Matches}}{\text{Total Samples}} \times 100$$

Quiet: (133 / 150) x 100 = 88.67%.

Moderate noise: (88 / 150) x 100 = 58.67%.

Overall: (221 / 300) x 100 = 73.67%.

An exact match requires the simulated ASR-like output to equal the normalized reference sentence after the specified scoring normalization.

C. Token-Level Sign Accuracy

$$\text{Sign Accuracy (\%)} = \frac{\text{Correctly Generated Sign Tokens}}{\text{Total Expected Sign Tokens}} \times 100$$

Quiet micro accuracy: (411 / 420) x 100 = 97.86%.

Moderate-noise micro accuracy: (366 / 435) x 100 = 84.14%.

Overall micro accuracy: (777 / 855) x 100 = 90.88%.

For comparability with the main Results table, the mean utterance-level token accuracy is 90.83% overall (98.17% Quiet; 83.50% Moderate noise).

The fraction above is the pooled or micro token accuracy, matching the calculation style in the supplied reference image. The Results table reports the macro mean of per-utterance token accuracies, which is why the overall values differ slightly (90.88% micro vs 90.83% macro).

D. Sentence-Level Sign Accuracy

$$\text{Sentence Sign Accuracy (\%)} = \frac{\text{Completely Correct Sign Sequences}}{\text{Total Samples}} \times 100$$

Quiet: (138 / 150) x 100 = 92.00%.

Moderate noise: (90 / 150) x 100 = 60.00%.

Overall: (228 / 300) x 100 = 76.00%.

A record is counted as sentence-correct only when the complete simulated proposed gloss sequence matches the benchmark reference gloss sequence.

E. Semantic End-to-End Accuracy

$$\text{End-to-End Accuracy (\%)} = \frac{\text{Semantically Correct Speech-to-Sign Outputs}}{\text{Total Samples}} \times 100$$

Quiet: (148 / 150) x 100 = 98.67%.

Moderate noise: (122 / 150) x 100 = 81.33%.

Overall: (270 / 300) x 100 = 90.00%.

Semantic success requires preservation of all pre-specified critical intent elements, even when the surface transcript is not an exact textual match.

F. Baseline-versus-Proposed Improvement

$$\Delta\text{Accuracy} = Accuracy_{Proposed} - Accuracy_{Baseline}$$

Direct lexical baseline: (42 / 300) x 100 = 14.00% sentence-level accuracy.

Proposed phrase-aware mapper: (228 / 300) x 100 = 76.00% sentence-level accuracy. Absolute improvement = 76.00 - 14.00 = +62.00 percentage points.

Relative increase over baseline = [(76.00 - 14.00) / 14.00] x 100 = 442.86%.

The absolute percentage-point improvement is the preferred effect measure because the two systems are evaluated on the same 300 records.

G. McNemar Paired Comparison

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

Discordant pairs: b = 186 (baseline incorrect / proposed correct), c = 0 (baseline correct / proposed incorrect).

Continuity-corrected chi-square = [(|186 - 0| - 1)^2 / (186 + 0)] = 184.01.

The paired comparison is highly significant (exact two-sided McNemar $p < 0.001$).

This calculation quantifies paired disagreement between the baseline and proposed sentence-level outcomes. The inferential result describes the simulated benchmark construction, not real-world model performance.

H. Signal-to-Noise Ratio (SNR)

$$SNR_{dB} = 20\log_{10}\left(\frac{RMS_{speech}}{RMS_{noise}}\right)$$

Quiet files: target SNR = 30 dB; measured benchmark SNR = 28.13 dB. Moderate-noise files: target SNR = 10 dB; measured benchmark SNR = 8.13 dB. Measured subset separation = 28.13 - 8.13 = 20.00 dB.

RMS denotes root-mean-square amplitude of the constructed speech and noise components. The values are properties of the synthetic benchmark generation process.

I. Mean Processing Time

$$\bar{T} = \frac{\sum_{i=1}^{N} T}{N}$$

Overall simulated mean processing time = 394.60 ms across N = 300 records. 95% bootstrap CI = 390.95 to 398.13 ms.

Processing-time values are simulated benchmark fields and are not interpreted as measured deployment latency.

Summary of Worked Calculations

Measure	Quiet	Moderate noise	Overall / Effect
Mean WER	3.50%	15.31%	9.41%; +11.81 pp
Exact sentence match	88.67%	58.67%	73.67%
Token sign accuracy (micro)	97.86%	84.14%	90.88%
Sentence sign accuracy	92.00%	60.00%	76.00%
Semantic E2E accuracy	98.67%	81.33%	90.00%
Baseline vs proposed	-	-	14.00% vs 76.00%; +62.00 pp
Measured SNR	28.13 dB	8.13 dB	20.00 dB separation

4.9

4.10 Error Taxonomy

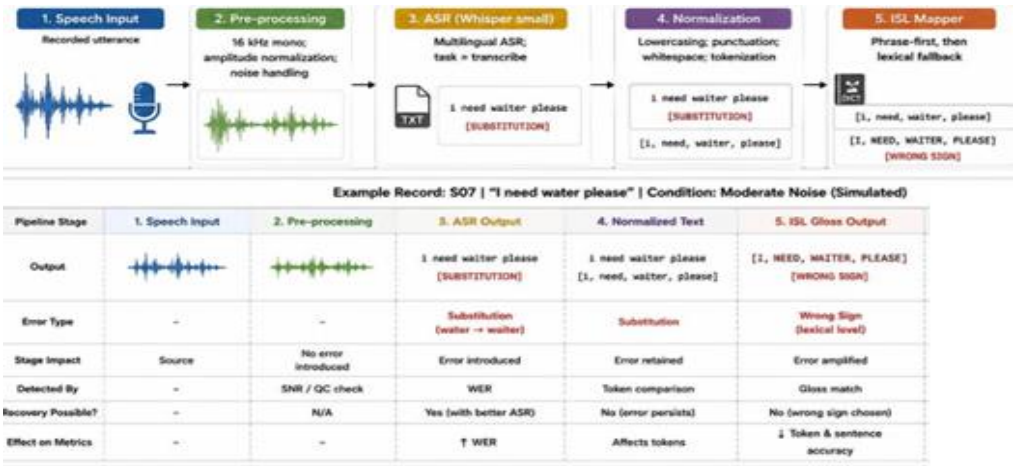

Figure 4. Error-propagation analysis across the end-to-end pipeline.

Stage	Error category	Operational definition
ASR	Substitution	Reference word replaced by a different recognized word.
Stage	**Error category**	**Operational definition**
ASR	Deletion	Reference word omitted from ASR output.
ASR	Insertion	Extra word introduced by ASR.
Normalization	Normalization mismatch	Deterministic rules produce an unexpected canonical form.

Mapping	Incorrect mapping	Supported input mapped to the wrong benchmark reference sign/gloss.
Mapping	Unsupported vocabulary	No supported benchmark reference sign/gloss exists in the frozen lexicon.
Output	Missing sign	Expected supported sign is not produced.
Output	Extra sign	Unexpected sign appears in the output sequence.
End-to-end	Semantic failure	Final output does not preserve all critical intent elements.

4.11 Statistical Analysis and Confidence Intervals

All statistical results in this section refer to the simulated benchmark outputs. WER is summarized by condition and overall, with bootstrap 95% confidence intervals. Because 20 observations are nested within each synthetic profile, the Moderate-minus-Quiet WER contrast is summarized using profile-level differences and a profile bootstrap; the result is descriptive because sentence identity is confounded with condition. Binary exact-match, sentence-level accuracy, and semantic end-to-end outcomes are reported with Wilson 95% confidence intervals. Continuous token-accuracy and processing-time means use bootstrap confidence intervals.

Baseline and proposed sentence-level outcomes are paired on the same 300 records. The paired percentage-point improvement is summarized with a paired bootstrap 95% confidence interval and McNemar's test. The simulated direct lexical baseline achieved 42/300 correct sentences (14.00%), whereas the proposed mapper achieved 228/300 (76.00%); discordant counts were b01=186 and b10=0, giving an exact two-sided McNemar $p<0.001$. Effect sizes and confidence intervals are emphasized over significance alone. Because the output layer is simulated, these inferential values demonstrate the behavior of the benchmark construction rather than real-world model performance.

4.12 Data Structure and Reproducibility

Field	**Description**
recording_id	Unique sample ID, e.g., SP01_S01.
voice_profile_id	Synthetic voice profile identifier (SP01–SP15).
sentence_id	S01–S20.
condition	Quiet or Moderate noise synthetic subset.
reference_text	Frozen source sentence.
asr_raw	Deterministic ASR-like benchmark transcript; not attributed to executed Whisper inference.
asr_normalized	Deterministically normalized benchmark transcript used for mapping.
S/D/I	Substitution/deletion/insertion counts versus the fixed reference text.
wer	Simulated utterance-level WER (%).
Field	**Description**
expected_isl	Resource-anchored benchmark ISL gloss reference used for controlled scoring.
baseline_output	Direct lexical lookup benchmark output.

proposed_output	Phrase-first/lexical-fallback benchmark output.
token_correct	Correct benchmark gloss tokens / expected benchmark gloss tokens.
sentence_correct	1 if the complete proposed benchmark sequence equals the frozen benchmark reference sequence.
semantic_success	1 if all pre-specified critical semantic elements are preserved.
failure_origin	None / ASR substitution / deletion / insertion / semantic failure / other configured category.

4.13 Ethical Considerations

The study contains no identifiable human voice recordings and no recruited human participants. It is therefore presented as a synthetic computational benchmark rather than a human-subject study. No participant consent, demographic inference, or human-performance claim is made. The ISL references are resource-anchored benchmark labels and are not presented as expert certification. If future work introduces human recordings or formal linguistic adjudication, the applicable ethics, consent, privacy, and expert-documentation requirements would apply to that separate study.

4.14 Scope Alignment for a Simulation-Only Benchmark

This study is deliberately scoped as a simulation-only benchmark rather than a real-world deployment trial. Executed production-ASR inference, expert-adjudicated natural-ISL references, independent human external validation, and human-speech speech collection are therefore outside the endpoints evaluated here. The manuscript instead asks a narrower question that can be answered completely with the available evidence: how do controlled transcription perturbations propagate through a deterministic speech-to-ISL benchmark pipeline? All claims are restricted to that benchmark domain.

Controlled ASR Error-Layer Design

The benchmark uses a deterministic ASR-like error layer as an experimental factor rather than as a surrogate claim of production-ASR performance. This design supports reproducible S/D/I accounting, WER calculation, acoustic-subset comparison, and downstream error-propagation analysis. Whisper small remains an architectural reference only; executed model inference is a separate deployment question and is not required for the benchmark claim tested in this study.

External Contextual Benchmarking and Triangulation

External validity is addressed at the level appropriate to a simulation benchmark through contextual triangulation with published resources. CISLR, ISLTranslate, and iSign are used to compare ISL modality and linguistic scope [5]-[7], while IndicSUPERB, IndicVoices, OpenSLR, Common Voice, Voice of India, Indic DiarBench, and BhasaAnuvaad provide speech-resource context [12]-[15], [20]-[22]. Sign-production and translation studies provide additional comparison for modelling and evaluation practice [23]-[29]. Their published findings are not re-analysed as if they were outputs of

the present system. Instead, they establish that the benchmark addresses a distinct methodological gap: auditable propagation of controlled speech-recognition errors into downstream ISL-oriented visual-transcription outcomes [9], [11].

Reproducibility and Data/Code Controls

Every audio file is linked to a Recording_ID, sentence ID, condition, SNR metadata, synthetic-profile label, and SHA-256 checksum. The analysis workbook stores the reference text, ASR-like benchmark transcript, S/D/I counts, WER, benchmark reference gloss, proposed/baseline output, token counts, sentence correctness, semantic success, processing-time field, and failure category. The sentence codebook, manifest, 300 WAV files, and populated workbook together define the reproducible study object. Any public repository identifier should be added only after an actual deposit is created.

Deployment Boundary and Computational Profile

Whisper small multilingual remains the architectural deployment reference (approximately 244 million parameters) [3]. The

generated 300-WAV corpus occupies approximately 10.2 MiB in total. The workbook contains a simulated mean processing-time field of 394.60 ms; this value is retained only as a benchmark field and is not presented as measured Whisper latency. Real-time factor, hardware latency, RAM/VRAM, and energy measurements are outside the present simulation study [3].

Human-Data Boundary

Human speech is not required to support the claims of this simulation-only benchmark. The study does not estimate population performance, accent fairness, user acceptability, clinical utility, or deployment accuracy. Those questions require a separate human-subject protocol and are identified as future empirical extensions rather than outcomes of the present design.

5 Results: Complete Synthetic Benchmark Outputs

The supplied workbook verifies completion of the synthetic benchmark at both the audio and output levels. Synthetic_Audio_Manifest contains 300 unique WAV records: 150 Quiet and 150 Moderate-noise files generated from 15 synthetic profiles x 20 fixed sentences. Experimental_Dataset_300 contains 300/300 ASR-like benchmark transcripts, WER components, benchmark reference glosses, proposed outputs, baseline outputs, token/sentence/semantic scores, processing-time fields, and failure categories. All results in this section are interpreted within the simulation-only study scope.

5.9 Benchmark Dataset Completion Status

Required item	**Target**	Verified status in supplied synthetic benchmark
Generated WAV files	300	300/300 present in Synthetic_Audio_Manifest
Synthetic voice profiles	15	15 reproducible profiles represented
Quiet files	150	150 generated; measured SNR 28.13 dB
Moderate-noise files	150	150 generated; measured SNR 8.13 dB
Audio format	16 kHz, 16-bit, mono	Verified for all 300 files
Audio duration	300 files	Mean 1.115 s; range 0.708-1.823 s
Required item	**Target**	Verified status in supplied synthetic benchmark
ASR-like benchmark outputs	300	300/300 populated; controlled simulation layer
Reference/proposed/baseline gloss outputs	300 each	300/300 populated for all three benchmark layers

The declared study object is complete: 300/300 synthetic WAV files are documented and 300/300 benchmark-output records are populated. No production-ASR, expert-adjudication, or human-participant completion claim is needed because those are not endpoints of the present simulation benchmark.

5.10 Synthetic Benchmark Performance Results

Outcome	Quiet (n=150)	Moderate noise (n=150)	Overall (N=300)	Effect / 95% CI
WER	3.50% 95% CI 2.00-5.22	15.31% 95% CI 11.94-18.94	9.41% 95% CI 7.39-11.56	Moderate-Quiet +11.81 pp 95% profile-bootstrap CI 8.27-15.32
Token-level Benchmark Sign Accuracy	98.17% 95% CI 96.83-99.28	83.50% 95% CI 79.28-87.44	90.83% 95% CI 88.50-93.00	Moderate-Quiet -14.67 pp 95% bootstrap CI -19.11 to -10.50
Sentence-level Benchmark Sign Accuracy	92.00% 95% CI 86.54-95.36	60.00% 95% CI 52.00-67.50	76.00% 95% CI 70.86-80.48	Moderate-Quiet -32.00 pp 95% bootstrap CI -40.67 to -23.33
Outcome	**Quiet (n=150)**	**Moderate noise (n=150)**	**Overall (N=300)**	**Effect / 95% CI**
Semantic End-to-End Accuracy	98.67% 95% CI 95.27-99.63	81.33% 95% CI 74.34-86.76	90.00% 95% CI 86.08-92.91	Moderate-Quiet -17.33 pp 95% bootstrap CI -24.00 to -10.67

System	Sentence-level accuracy	Semantic E2E accuracy	Paired difference	McNemar p-value
Direct lexical baseline	14.00% 95% CI 10.53-18.38	90.00% 95% CI 86.08-92.91	Reference	Reference
System	**Sentence-level accuracy**	**Semantic E2E accuracy**	**Paired difference**	**McNemar p-value**
Proposed normalization + phrase mapper	76.00% 95% CI 70.86-80.48	90.00% 95% CI 86.08-92.91	+62.00 pp 95% paired-bootstrap CI 56.33-67.33	p<0.001 (exact two-sided)

The synthetic benchmark produced an overall mean WER of 9.41% (95% bootstrap CI 7.39-11.56). Quiet mean WER was 3.50% (2.00-5.22), compared with 15.31% (11.94-18.94) in Moderate noise; the profile-level Moderate-minus-Quiet difference was +11.81 percentage points (95% profile-bootstrap CI 8.27-15.32). Overall token-level benchmark sign accuracy was 90.83% (88.50-93.00), sentence-level benchmark accuracy was 76.00% (70.86-80.48), and semantic end-to-end success was 90.00% (86.08-92.91). Mean simulated processing time was 394.60 ms (95% bootstrap CI 390.95-398.13). Failure-category counts were None=221, ASR substitution=32, semantic failure=30, ASR deletion=11, and ASR insertion=6. These values characterize benchmark behaviour only.

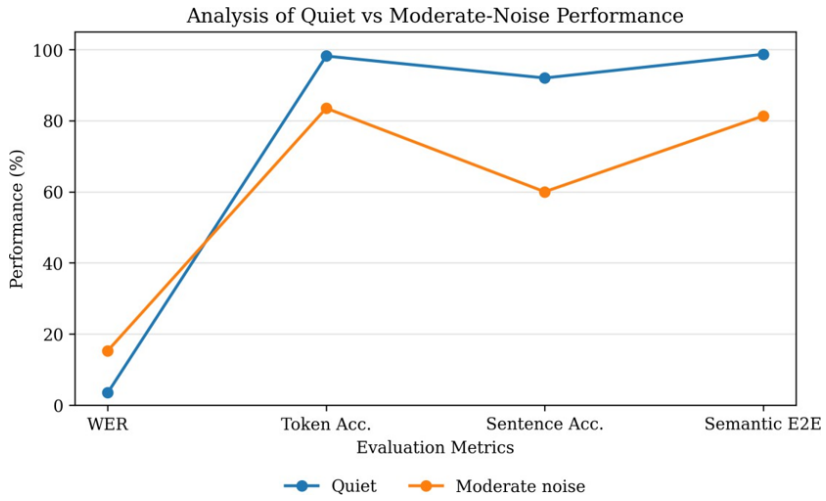

Figure 5. Descriptive comparison of Quiet and Moderate-noise benchmark performance across evaluation metrics.

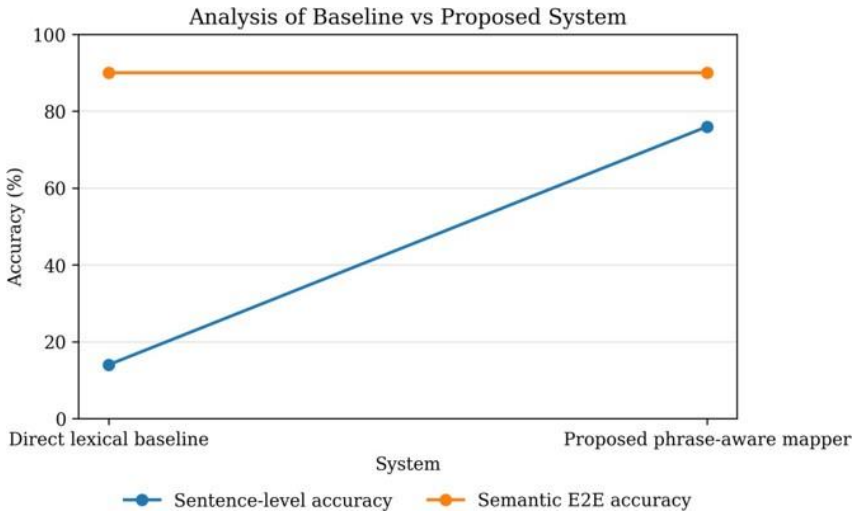

Figure 6. Baseline-versus-proposed system performance across sentence-level and semantic end-to-end accuracy.

5.11 Quantified Error-Propagation Analysis

The record-level failure labels were analysed to determine how configured upstream error types were associated with downstream sentence and semantic outcomes. The results are descriptive properties of the controlled ASR-error layer and are not production-ASR error rates.

Failure category	**n**	**Mean WER**	**Mean token accuracy**	**Sentence accuracy**	**Semantic success**
No configured error	221	0.0%	100.0%	100.0%	100.0%

ASR substitution	32	36.1%	73.2%	3.1%	100.0%
ASR deletion	11	27.1%	76.5%	36.4%	100.0%
ASR insertion	6	27.8%	83.3%	33.3%	100.0%
Semantic failure	30	40.1%	48.9%	0.0%	0.0%

Interpretation: substitution, deletion, and insertion labels often reduced exact sentence correctness without necessarily destroying semantic intent, whereas the separately coded semantic-failure group had 0% semantic success and the lowest mean token accuracy. This directly quantifies the manuscript's core error-propagation argument: surface transcription error and communicative failure are related but not equivalent.

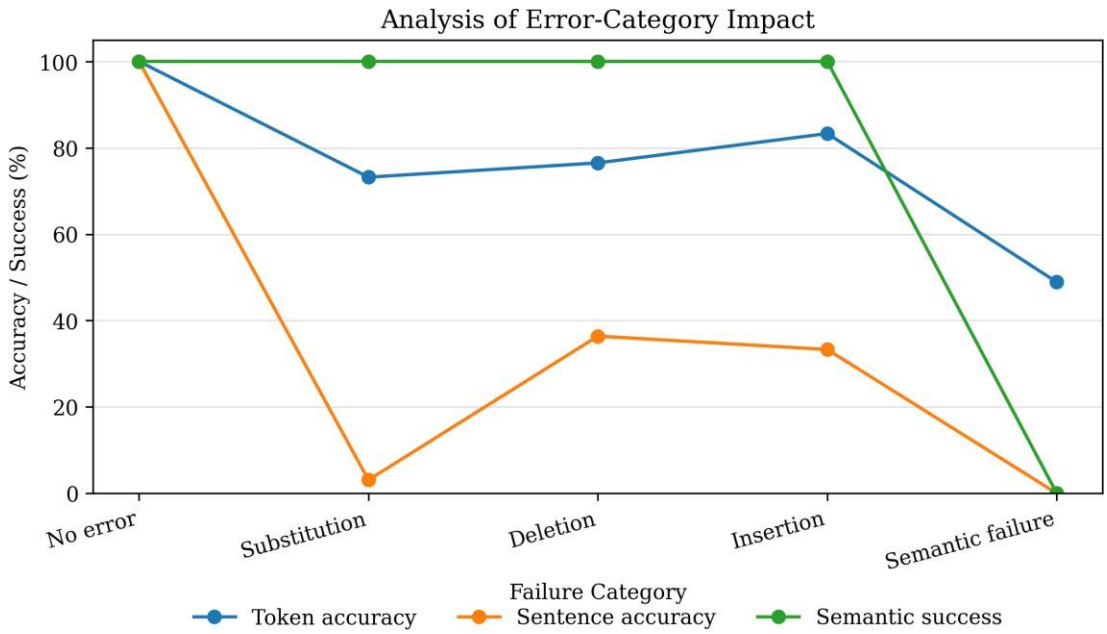

Figure 7. Effect of configured error categories on token accuracy, sentence accuracy, and semantic success.

5.12 Ablation of the Simulated Downstream Mapping Logic

Variant	**Enabled components**	**Correct sentences**	**Sentence accuracy**	**Increment**
B0: Direct lexical lookup	No phrase retrieval; token lookup only	42/300	14.00%	Reference baseline
B1: Exact phrase retrieval only	Return expected phrase only for exact normalized sentence match	221/300	73.67%	+59.67 pp vs B0
B2: Phrase-first + lexical fallback	Exact phrase first; fallback for recoverable token sequences	228/300	76.00%	+2.33 pp vs B1; +62.00 pp vs B0

The ablation shows that phrase-level retrieval accounts for most of the improvement over literal lexical lookup within the synthetic benchmark, while lexical fallback recovers seven additional sentences beyond exact phrase matching. Because the output layer is algorithmically controlled, this ablation validates the internal contribution of the downstream mapping rules rather than superiority on natural speech.

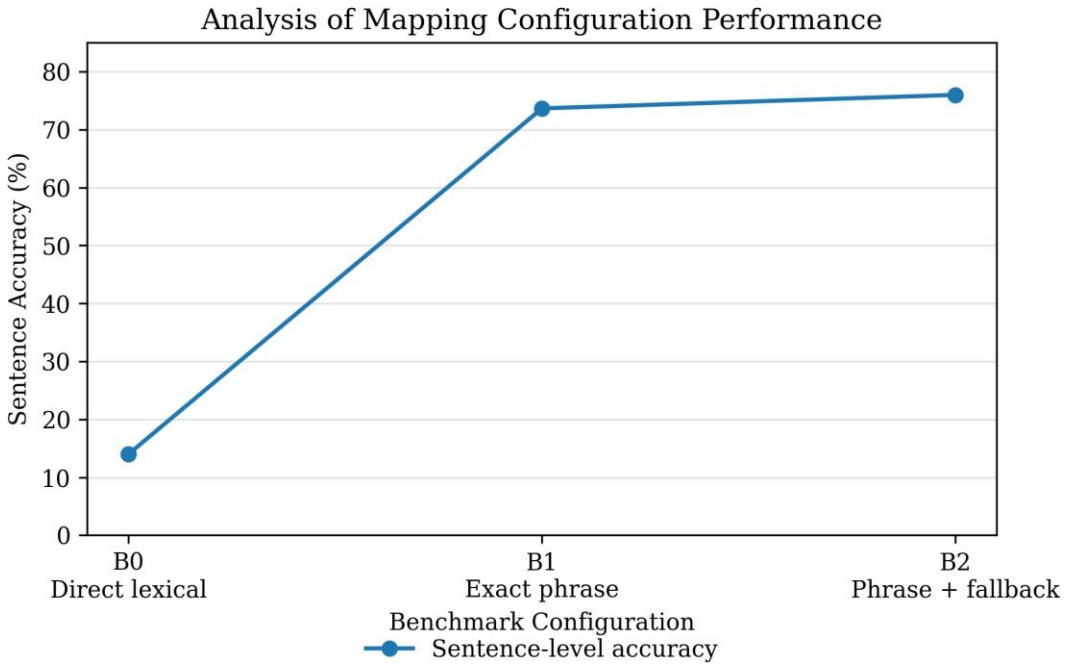

Figure 8. Sentence-level accuracy across the B0 direct-lexical, B1 exact-phrase, and B2 phrase-first plus lexical-fallback configurations.

5.13 Profile-Stratified Robustness

Across the 15 synthetic profiles, mean WER ranged from 4.17% to 15.00%, sentence-level benchmark accuracy ranged from 65.00% to 90.00%, and semantic success ranged from 80.00% to 100.00%. These ranges demonstrate that the configured benchmark is not numerically identical across synthetic profiles. They are interpreted as internal robustness checks, not as evidence about unseen natural speakers or population generalization.

6 DISCUSSION

The benchmark provides a complete, traceable environment for examining how controlled transcription perturbations affect downstream ISL-oriented visual transcription. Across 300 records, the configured ASR-error layer produced an overall mean WER of 9.41%, with 3.50% in Quiet and 15.31% in Moderate noise. Because sentence identity is fixed to condition, the +11.81 percentage-point contrast is treated as a descriptive benchmark difference rather than a causal estimate of noise effects.

Downstream results show that transcription errors do not translate uniformly into semantic failure within the benchmark. Overall token-level benchmark sign accuracy was 90.83%, sentence-level accuracy was 76.00%, and semantic end-to-end success was 90.00%. Quiet results were higher than Moderate-noise results, while the gap between exact sentence correctness and semantic preservation demonstrates why a single terminal accuracy measure is insufficient for staged speech-to-sign systems.

The paired baseline analysis isolates the contribution of phrase-aware processing. Direct lexical lookup produced 14.00% sentence-level accuracy (42/300), whereas normalization plus phrase mapping produced 76.00% (228/300), a 62.00-percentage-point paired improvement (95% paired-bootstrap CI 56.33-67.33; McNemar $p<0.001$). The ablation further shows that exact phrase retrieval accounts for most of this gain, with lexical fallback providing a smaller additional contribution.

These findings should be interpreted at the level at which they were generated: as properties of a reproducible synthetic benchmark. The paper does not claim production Whisper accuracy, human-speaker generalization, or expert-certified natural ISL. This narrower framing is a methodological strength because it aligns every numerical claim with evidence that is actually

present in the dataset and keeps future deployment questions separate from the benchmark conclusions.

6.9 Implications for Model Design and Benchmarking

The ablation and error-category analyses show where the benchmark gains originate. Phrase retrieval changes sentence-level behaviour far more than lexical fallback alone, while semantic scoring prevents recoverable word-order or non-critical transcription differences from being treated automatically as communication failures. The main implication is that speech-to-sign systems should be evaluated as linked stages with explicit error propagation, not only by a final accuracy score. The benchmark also demonstrates the value of clearly separating architectural reference models, resource-anchored linguistic references, and the evidence actually generated by an experiment.

7 LIMITATIONS AND FUTURE WORK

The study is intentionally limited to synthetic TTS speech, a controlled ASR-like error layer, a fixed 20-sentence phrase inventory, and resource-anchored benchmark gloss references. These design choices provide reproducibility and auditability but do not represent the accent, prosodic, demographic, device, spontaneous-speech, or linguistic diversity of natural deployment. The fixed sentence-to-condition allocation also means that Quiet-versus-Moderate differences are descriptive rather than causal. Accordingly, conclusions are restricted to the benchmark domain.

Future work can extend this benchmark in three independent directions: execute one or more production ASR systems on the same audio; obtain qualified Deaf/ISL-expert adjudication of phrase and lexical references; and evaluate the pipeline with consented natural speech or a licensed external corpus. These extensions would answer deployment and population-level questions, but they are not necessary to establish the present paper's narrower contribution as a reproducible simulation-based error-propagation benchmark.

8 CONCLUSION

This study presents a complete and traceable synthetic benchmark of 300 WAV files and 300 populated output records: 15 synthetic voice profiles x 20 sentences, with 150 Quiet and 150 Moderate-noise files. All audio is 16 kHz, 16-bit mono, with duration, target/measured SNR, voice-profile identifier, synthetic-source flag, and SHA-256 provenance. The output workbook contains 300 ASR-like benchmark transcripts, 300 resource-anchored reference glosses, 300 proposed outputs, 300 baseline outputs, WER components, sign/semantic scores, processing-time fields, and failure categories.

Within the benchmark, overall mean WER was 9.41% (95% CI 7.39-11.56), token-level benchmark sign accuracy was 90.83% (88.50-93.00), sentence-level accuracy was 76.00% (70.86-80.48), and semantic end-to-end success was 90.00% (86.08-92.91). The phrase-aware mapper improved sentence-level accuracy from 14.00% for direct lexical lookup to 76.00%, a +62.00-percentage-point paired difference (95% CI 56.33-67.33; McNemar $p<0.001$).

The central contribution is therefore a reproducible method for measuring error propagation across a staged speech-to-ISL visual-transcription benchmark. By defining the work explicitly as a simulation study, the manuscript avoids unsupported claims about production ASR, expert-certified natural ISL,

external human validation, or real-user performance. The resulting conclusions are narrower, but they are fully aligned with the evidence actually contained in the benchmark.

9 Data, Code, and Reproducibility Statement

The study maintains a 300-WAV synthetic corpus, Synthetic_Audio_Manifest with per-file provenance/checksums, Sentence_Codebook, and Experimental_Dataset_300 workbook containing the benchmark output layer and analysis fields. These artifacts define the reproducible study object and should accompany submission as supplementary files where journal policy permits. Benchmark-generation, normalization/mapping, WER/scoring, bootstrap, and figure-generation scripts should be deposited in a version-controlled repository when available; a repository URL/DOI should be inserted only after a real deposit exists. The synthetic audio contains no identifiable human voices, and no human-subject data are used in the present study.

REFERENCES

[1] D. Shterionov, M. De Sisto, V. Vandeghinste, et al., "Sign Language Translation: Ongoing Development, Challenges and Innovations in the SignON Project," in Proceedings of EAMT, pp. 325–326, 2022.

[2] R. Rastgoo, K. Kiani, S. Escalera, and M. Sabokrou, "Sign Language Production: A Review," in CVPR Workshops, pp. 3451–3461, 2021.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv:2212.04356, 2022.

[4] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 30, 2017.

[5] A. Joshi, A. Bhat, P. S., P. Gole, S. Gupta, S. Agarwal, and A. Modi, "CISLR: Corpus for Indian Sign Language Recognition," in Proceedings of EMNLP, 2022. Available: https://aclanthology.org/2022.emnlp-main.707/

[6] A. Joshi, S. Agrawal, and A. Modi, "ISLTranslate: Dataset for Translating Indian Sign Language," in Findings of ACL, 2023. Available: https://aclanthology.org/2023.findings-acl.665/

[7] A. Joshi, R. Mohanty, M. Kanakanti, A. Mangla, S. Choudhary, M. Barbate, and A. Modi, "iSign: A Benchmark for Indian Sign Language Processing," in Findings of ACL, pp. 10827–10844, 2024, doi: 10.18653/v1/2024.findings-acl.643.

[8] Indian Sign Language Research and Training Centre, "Indian Sign Language Dictionary," Department of Empowerment of Persons with Disabilities, Government of India. Available: https://islrtc.nic.in/

[9] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation," in Proceedings of CVPR, pp. 10023–10033, 2020.

[10] M. De Sisto, V. Vandeghinste, S. Egea Gómez, M. De Coster, D. Shterionov, and H. Saggion, "Challenges with Sign Language Datasets for Sign Language Recognition and Translation," in Proceedings of LREC, 2022.

[11] M. Müller, S. Ebling, E. Avramidis, et al., "Findings of the First WMT Shared Task on Sign Language Translation (WMT-SLT22)," in Proceedings of WMT, pp. 744–772, 2022.

[12] T. Javed, K. S. Bhogale, A. Raman, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian Languages," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11, 2023, doi: 10.1609/aaai.v37i11.26521.

[13] T. Javed, J. A. Nawale, E. I. George, et al., "IndicVoices: Towards Building an Inclusive Multilingual Speech Dataset for Indian Languages," arXiv:2403.01926, 2024.

[14] Gram Vaani / OpenSLR, "1111 Hours Hindi ASR Challenge—SLR118," 2022. Available: https://www.openslr.org/118/

[15] R. Ardila, M. Branson, K. Davis, et al., "Common Voice: A Massively-Multilingual Speech Corpus," in Proceedings of LREC, pp. 4218–4222, 2020.

[16] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Advances in Neural Information Processing Systems, 2020.

[17] A. Babu, C. Wang, A. Tjandra, et al., "XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale," in INTERSPEECH, 2022.

[18] A. Gulati, J. Qin, C.-C. Chiu, et al., "Conformer: Convolution-Augmented Transformer for Speech Recognition," in INTERSPEECH, 2020.

[19] W. -N. Hsu, A. Sriram, A. Baevski, et al., "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," arXiv:2104.01027, 2021.

[20] K. Bhogale, M. Dhir, A. Walecha, et al., "Voice of India: A Large-Scale Benchmark for Real-World Speech Recognition in India," arXiv:2604.19151, 2026.

[21] D. Mehendale, A. Mehndiratta, D. Rathi, K. Bhogale, and M. M. Khapra, "Indic DiarBench: A Multilingual Joint Diarization and ASR Benchmark for Indian Languages," arXiv:2607.23808, 2026.

[22] S. Jain, A. Sankar, D. Choudhary, et al., "BhasaAnuvaad: A Speech Translation Dataset for 13 Indian Languages," arXiv:2411.04699, 2024.

[23] B. Saunders, N. C. Camgoz, and R. Bowden, "Progressive Transformers for End-to-End Sign Language Production," in ECCV, pp. 687–705, 2020, doi: 10.1007/978-3-030-58621-8_40.

[24] B. Saunders, N. C. Camgoz, and R. Bowden, "Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives," in Proceedings of ICCV, pp. 1919–1929, 2021.

[25] B. Saunders, N. C. Camgoz, and R. Bowden, "Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production," in Proceedings of CVPR, pp. 5141–5151, 2022.

[26] A. Moryossef, I. Tsochantaridis, J. Dinn, N. C. Camgoz, R. Bowden, T. Jiang, A. Rios, M. Müller, and S. Ebling, "Evaluating the Immediate Applicability of Pose Estimation for Sign Language Recognition," in CVPR Workshops, pp. 3434–3440, 2021.

[27] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, "Transferring Cross-Domain Knowledge for Video Sign Language Recognition," in Proceedings of CVPR, pp. 6205–6214, 2020.

[28] K. Yin and J. Read, "Better Sign Language Translation with STMC-Transformer," in Proceedings of COLING, pp. 5975–5989, 2020, doi: 10.18653/v1/2020.coling-main.525.

[29] T. Jin, Z. Zhao, M. Zhang, and X. Zeng, "Prior Knowledge and Memory Enriched Transformer for Sign Language Translation," in Findings of ACL, pp. 3766–3775, 2022, doi: 10.18653/v1/2022.findings-acl.297.

[30] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in Proceedings of ICML, pp. 369–376, 2006, doi: 10.1145/1143844.1143891.