From Script to Speech: Contrast-Preserving Grapheme-to-Phoneme Frontends and Text-to-Speech for Six Languages of Nepal

Ampixa
https://ampixa.com

***Abstract*—High-resource text-to-speech can learn from characters or bytes because large paired corpora expose the spelling–sound relation many times. For under-resourced languages that share a dominant script, the first requirement is to preserve the contrasts that make words distinct. We present a family of deterministic G2P frontends, along with accompanying TTS models, for six languages of Nepal: five severely under-resourced languages — Limbu, Magar, Bantawa, Gurung, and Dhimal — along with Nepali, the high-resource anchor. These under-resourced languages extend Devanagari with extra signs, nukta vowels, visarga length, glottal letters, lexical vowel contrasts, and other elements. A generic Nepali frontend silently erases these elements, collapsing glottal stop, breathy voice, contrastive vowel length, voiceless sonorants, gemination, and other features. We introduce a contrast-preservation evaluation that needs no full pronunciation dictionary, use it to audit text frontends before synthesis, and then carry those sets to TTS diagnostics. At the G2P level, our frontends preserve marked contrasts where Nepali baselines score 0%: 96.7% for Limbu, 100% for Magar, and 100% for Gurung. Dhimal reaches 96.4% lexicon exact match. For hidden contrasts, Bantawa recovers /ɨ/ lexically for 36.9% of attested words. To the best of our knowledge, Magar, Bantawa, and Gurung have no released language-specific TTS or G2P system beyond MMS language-identification coverage. Limbu and Dhimal have MMS checkpoints but no explicit frontend. We use the frontends to drive TTS and report contrast-specific diagnostics, including a Limbu vowel-length win over MMS. We release the frontends, tooling, and TTS model checkpoints.**

***Index Terms*—grapheme-to-phoneme, text-to-speech, low-resource languages, Devanagari, Nepal**

I. Introduction

Grapheme-to-phoneme conversion is a small component, but the consequences are large. If a writing system does not map transparently to the phoneme inventory of the language being spoken, the frontend can remove word-level contrasts before a recogniser or synthesiser sees them. This is routine in high-resource languages, where the frontend is treated as infrastructure. Most languages do not have that infrastructure [1], [2].

This paper studies that problem for six languages of Nepal written in Devanagari: Limbu, Magar, Bantawa, Gurung, Dhimal, and Nepali. Nepali is not the novelty target. It is the high-resource anchor and the baseline that a practitioner would naturally reach for. The other five apply the same script to phonologies that Nepali does not have: glottal stop, vowel length, lower-mid vowels, breathy voice, voiceless or murmured sonorants, and the high central vowel /ɨ/. A Nepali frontend therefore does not merely sound foreign on these languages; it collapses distinctions that identify words.

We make four contributions. First, we provide deterministic, inspectable Devanagari G2P frontends for all six languages, grounded in published grammars and local lexical resources. Second, we describe a gold-free contrast-preservation test for cases where the orthography overtly marks a contrast: Limbu, Magar, and Gurung score 96.7%, 100%, and 100%, while two Nepali baselines score 0%. Third, we separate that from the harder recovery case where the orthography hides the contrast: Bantawa recovers /ɨ/ lexically for 36.9% of attested /ɨ/ words, and Dhimal matches an independent lexicon at 96.4% exact word accuracy. Fourth, we use the frontends to drive TTS and report both wins and losses: in Limbu, the explicit frontend realises vowel length where MMS is at chance, but MMS remains cleaner overall; in Magar, an automatic breathy-voice probe is a null result. We do not claim TTS superiority.

The novelty claim is intentionally layered. Magar, Bantawa, and Gurung are covered by MMS for spoken language identification, but not for released MMS TTS or MMS ASR; to the best of our search, no released language-specific TTS, ASR, or G2P system exists for those three languages beyond LID coverage. Our claim is therefore first TTS/G2P for Magar, Bantawa, and Gurung, not first speech technology of any kind. Dhimal already has MMS-TTS, so our claim is a first explicit standalone G2P frontend. Limbu has MMS-TTS and a small WikiPron list, so the claim is first large-scale Devanagari G2P rather than first speech technology.

II. Languages and Contrasts

The common engineering problem is that Devanagari is being used as a shared script for languages whose phonological inventories are not shared. Table II gives the contrasts used by the frontends and the evaluation source for each language.

The contrast inventories are grounded in published grammars and orthography descriptions [3], [4], [5], [6], [7], [8], [9], then checked against the language-specific lexica and speech corpora available to the project.

These languages expose two distinct script–phonology mismatches. In the first, the orthography marks a contrast, but a generic Nepali frontend ignores or breaks it. Limbu writes the glottal stop as ʔ, length as ः, and lower-mid vowels with nukta; Magar writes final क where the target phone is /ʔ/ and uses conjuncts for breathy series; Gurung writes sonorant+ ्ह

TABLE I

POSITIONING AGAINST AVAILABLE ALTERNATIVES. "BETTER" HERE MEANS BETTER FOR THE PAPER'S TARGETED TASK: PRESERVING OR RECOVERING THE NON-NEPALI CONTRASTS IN DEVANAGARI TEXT WITH AN INSPECTABLE FRONTEND. IT IS NOT A CLAIM OF GLOBALLY BETTER ACOUSTIC QUALITY.

System or resource	Explicit G2P	Target-lang. coverage	Contrast test?	What it provides	Why our frontends differ
MMS character speech models [16]	no	Limbu, Dhimal TTS/ASR; Magar, Bantawa, Gurung LID only	no	neural speech checkpoints and broad LID coverage	inspectable phone decisions; first released TTS/G2P path for Magar, Bantawa, Gurung; Limbu length contrast beats MMS despite lower overall quality
Nepali G2P and espeak-ng Nepali	yes, for Nepali	Nepali only	yes, as baselines	realistic fallback for Devanagari text	score 0% on marked Limbu, Magar, Gurung contrasts; ours scores 96.7/100/100%
Epitran / Indic script tools [10]	yes, broad rules	major supported languages/scripts	no	useful multilingual IPA or script conversion where language modules exist	no language-specific modules for these Devanagari adaptations; does not encode the local glottal, length, breathy, or voiceless-sonorant decisions
WikiPron and multilingual neural G2P [11], [12]	mined or modelled	sparse or absent for targets	no	pronunciation mining and generic multilingual G2P models	no large Devanagari frontend or contrast audit for these languages; ours is grammar/lexicon-grounded and reports contrast-level statistics

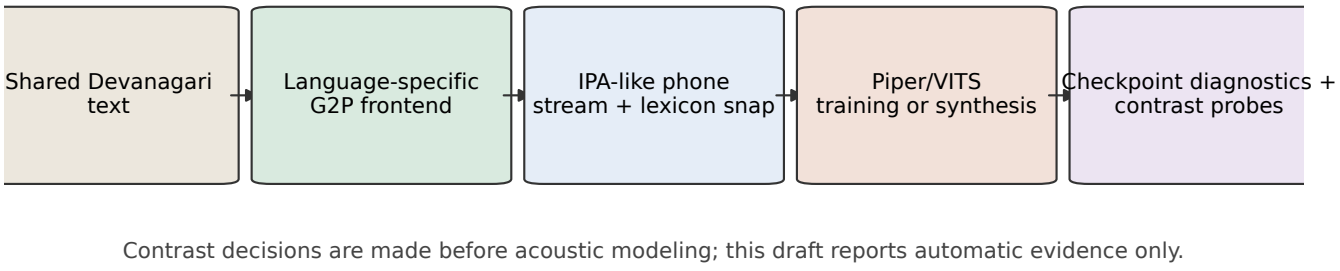


Fig. 1. Paper pipeline. Language-specific G2P is the inspectable decision point; TTS and automatic diagnostics evaluate how much of that contrast reaches speech.

clusters that represent single voiceless sonorants. In the second mismatch, the orthography hides a contrast. Bantawa writes /ɨ/ mostly as the inherent vowel, and Dhimal's New Testament orthography does not mark the glottal stop or length contrasts described by King. The latter case is still a frontend problem, but it is not the same metric.

III. G2P FRAMEWORK

All frontends are deterministic. They normalise Unicode, treat Devanagari as akshara-like units, emit a broad phoneme stream, and apply language-specific rules for inherent vowels, virama, diacritics, conjuncts, and lexical overrides. The design choice is conservative. Every special case is inspectable and can be traced to a grammar, lexicon, corpus statistic, or later native-speaker review.

Limbu illustrates the marked-contrast case. The frontend maps ʔ to /ʔ/, ः to length /ː/, and nukta-marked vowel signs to /ɛ/ and /ɔ/. It does not apply Nepali schwa deletion; its source corpus overwhelmingly gives unambiguous word endings, and remaining bare endings are mostly loans or names. A Nepali frontend cannot represent the resulting inventory, so it drops the mark or maps it to the wrong segment.

Bantawa illustrates the hidden-contrast case. Doornenbal's lexicon shows frequent /ɨ/, but the Devanagari corpus writes that vowel predominantly with the inherent vowel: in the New Testament corpus, 4,942 of 5,196 lexicon-snapped /ɨ/ positions are written as *a*. The Bantawa frontend therefore combines rules with a conservative lexicon-snap layer. If a rule-generated consonant skeleton uniquely identifies a Doornenbal headword modulo the hidden vowel and glottal alternations, the output snaps to the lexicon form; otherwise it falls back to the rule output rather than guessing. On the TTS corpus, the snap layer applies to 33,297 of 123,916 running word tokens (26.9%) and changes the emitted phone string for 5,387 tokens (4.3%), mainly recovering hidden /ɨ/ and /ʔ/. This cannot be treated as a 100% spelling rule, and the paper reports it as lexical recovery.

Dhimal is similar in spirit but has a stronger lexicon. After documented convention normalisation, the frontend reaches 96.4% exact match against 7,058 SIL Webonary pronunciations and 0.59% character PER, while a corpus gate over 7,830 New Testament lines emits 986,448 phonemes with no unmapped graphemes and no non-IPA leakage. King 2009 independently confirms the relevant contrasts: a glossary cross-

TABLE II

LANGUAGES, COMMITTED FRONTEND CONTRASTS, AND EVIDENCE. "MARKED" CONTRASTS ARE OVERTLY RECOVERABLE FROM DEVANAGARI AND ENTER THE PRESERVATION TEST; "HIDDEN" CONTRASTS REQUIRE LEXICAL OR AUDIO EVIDENCE AND ARE REPORTED SEPARATELY.

Language	ISO	Family/use in paper	Frontend contrast	Primary grounding
Limbu	lif	Kiranti; worked example	ʔ→/ʔ, ः→/ː, nukta vowels /ɛ ɔ/	van Driem; Michailovsky & Everson; Noonan
Magar	mgp	Tibeto-Burman; first TTS/G2P	final/coda क→/ʔ; ◌्ह breathy series	SIL Western Magar dictionary; Grunow-Hårsta
Gurung	gvr	Tamangic; first TTS/G2P	ल्ह म्ह न्ह ङ्ह→/ɬ m̥ n̥ ŋ̥	Glover & Glover; Tamukyi app; audio pipeline
Bantawa	bap	Kiranti; first TTS/G2P	hidden /ɨ/ and /ʔ/ via lexicon snap	Doornenbal grammar and IPA lexicon
Dhimal	dhi	Dhimalic; explicit-G2P claim	hidden glottal/length; schwa and breathy series	King grammar; SIL Webonary lexicon
Nepali	npi	baseline anchor	no target novelty	existing production frontend

TABLE III

FLAVOR A: ORTHOGRAPHY MARKS THE CONTRAST; THE TEST ASKS WHETHER THE CONTRAST SURVIVES THE FRONTEND.

Language	contrast set	ours	95% CI	Nepali	espeak
Limbu	glottal, length, lower-mid vowels	**580/600**	[94.9,97.8]	0/600	0/600
Magar	final क→ʔ; ◌्ह breathy series	**284/284**	[98.7,100]	0/284	0/284
Gurung	ल्ह म्ह न्ह ङ्ह sonorants	**761/761**	[99.5,100]	0/761	0/761

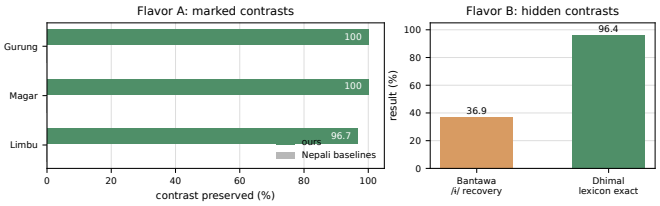


Fig. 2. Current G2P results, split by the two evaluation flavors. Marked-contrast preservation is not pooled with hidden-contrast lexical recovery.

check gives 847 byte-identical King–SIL agreements, including source evidence for 320 glottal-marked and 167 length-marked words. The NT spelling omits glottal stop and length, so those contrasts require audio or native review, not more text rules.

IV. CONTRAST EVALUATION

The evaluation framework is designed around what the orthography makes observable. It has two flavors because a single aggregate score would hide two different failure modes. Flavor A covers the three languages where Devanagari overtly marks a non-Nepali contrast: Limbu, Magar, and Gurung. This test needs no trusted full pronunciation dictionary. We sample attested word types containing the mark, compare a system's output for the word against its output after the mark is removed, and classify the result as *correct* (the committed contrastive feature appears), *collapsed* (outputs are identical or the contrast is lost), or *corrupted* (the strings differ, but by the wrong feature). The comparison systems are realistic baselines for a user trying Devanagari text: the production Nepali frontend and espeak-ng Nepali. This setup avoids falsely crediting a system that makes strings different by spelling out a mark or inserting /h/.

For both flavors, the unit of analysis is a word type rather than a token, so highly repeated Bible or curriculum words do not dominate the estimate. We report exact counts, percentages, and Wilson 95% confidence intervals for binomial-style recovery or preservation rates. The 100% rows are therefore finite-sample results with lower confidence bounds, not claims of universal correctness.

TABLE IV

FLAVOR B: ORTHOGRAPHY HIDES THE CONTRAST; THE FRONTEND MUST RECOVER IT LEXICALLY. THESE ARE DIFFERENT METRICS AND ARE NOT POOLED WITH TABLE III.

Language	metric	result	statistic
Bantawa	/ɨ/ recovery	**83/225**	36.9%, CI [30.9,43.4]
Dhimal	lexicon exact match	**6804/7058**	96.4%, CI [95.9,96.8]; char-PER 0.59%

Table III is the headline preservation result. Limbu's 600-word evaluation samples 200 attested words for each mark; the sub-scores are 100% for glottal stop, 90.5% for length, and 99.5% for lower-mid vowels. The length misses are a documented decomposed-vowel edge case, not a loss of the length segment. Magar's result covers 141 final-क cases and 143 ◌्ह conjunct cases, and also has an independent coarse PER comparison against the SIL dictionary: 11.0% for the Magar frontend versus 94.4% for the production Nepali G2P and 65.5% for espeak-ng. Gurung's 761 cases cover four voiceless-sonorant clusters. All three preservation results are stable across ten random seeds, and the Nepali baselines are 0% on every seed because their phone inventories lack the target contrasts.

Flavor B covers languages whose corpus orthography hides the contrast. Here the mark-drop test would be meaningless because there is no mark to drop; the evaluation must be lexical. Bantawa has no clean Devanagari mark for glottal stop or vowel length in the available corpus, and /ɨ/ is written as

TABLE V
CURRENT TTS EVIDENCE AND LIMITS FOR THE FIRST REVISION.

Lang.	status	strongest speech number	caveat
Limbu	ours vs MMS	length median 1.59× vs 1.09×; DNSMOS loses 2.90 vs 3.22	only length is a win
Bantawa	epoch 59	MCD-DTW 6.96 dB; ECAPA 0.696; PER 0.591	no competitor TTS
Magar	epoch 49; breathy probe epoch 29	MCD 9.846; PER 0.589; DNSMOS 2.937; SECS 0.812	breathy contrast null
Gurung	epoch 19	MCD 13.98→8.85→8.87; /ɬ/ probe $p = 0.042$	nasals null
Dhimal	manifests ready	7,844 kept verses; 25,887 clips; 50 phones	no audited checkpoint

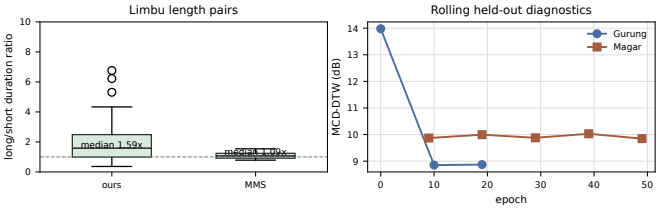


Fig. 3. Speech diagnostics available for the first revision. Left: Limbu long/short vowel ratios show the narrow contrast-specific win over MMS. Right: rolling MCD-DTW curves for Magar and Gurung are checkpoint diagnostics, not final quality claims.

ordinary *a*. The 36.9% number therefore measures lexical recovery on 225 attested /ɬ/ words, not orthographic preservation. The 142 misses are ambiguous skeletons where the snap layer refuses to guess, so they fall back to the Nepali-like reading rather than producing an incorrect special phone. Dhimal's number is likewise lexicon-grounded segmental accuracy, not a contrast percentage: the NT spelling does not mark the glottal and length contrasts that the grammar records.

V. From Text to Speech

The frontends feed Piper/VITS-style single-speaker TTS pipelines, with natural reference audio drawn from Bible recordings where available [14], [15]. The speech evaluation is designed in two layers. First, each checkpoint receives diagnostics that do not require a competing system: held-out MCD-DTW, Allosaurus phone-PER, DNSMOS, and speaker similarity where available. Second, when a contrast has an automatic acoustic proxy, we test it directly: duration ratio and a paired sign test for vowel length, H1–H2 and a one-sided Mann–Whitney test for breathy voice, and onset voicing fractions for Gurung's voiceless sonorants. Automatic signal quality by itself is not the contribution. The contribution is whether a language-aware frontend can put a contrast into the input representation and, where the acoustic model cooperates, into the waveform.

The comparison regime depends on the released baseline. MMS gives a direct Limbu comparison now and a Dhimal comparison once our checkpoint is audited. For Magar, Bantawa, and Gurung there is no released MMS TTS/ASR comparator, so those rows are framed as first-system diagnostics rather than head-to-head wins.

Limbu is the deepest completed comparison because Meta MMS provides a Limbu TTS voice. Both systems are VITS-like, but MMS is a mature character-level model while ours is an early explicit-G2P checkpoint, so overall quality is confounded. The contrast-specific result is narrow but strong: on 36 attested monosyllabic length minimal pairs, our system gives a median long/short voiced duration ratio of 1.59× (mean 2.99×), with the long member longer in 75% of pairs (sign-test $p = 0.004$). MMS is at chance: median 1.09×, 58%, $p = 0.41$. This is the one automatic TTS win we claim.

The losses matter. MMS is cleaner by DNSMOS (OVRL 3.22 versus 2.90) and has lower Allosaurus phone-PER on 16 held-out sentences (0.444 versus 0.551) [17], [18], even though that PER reference is biased toward our own frontend. Automatic glottal-stop probes are null for both systems: creak is uncorrelated with ʔ count, and a weak closure correlation disappears after controlling for voiceless obstruents. Gemination is not demonstrable. So the speech claim is not that the new voice is better than MMS. It is that explicit G2P gives one robust phonemic contrast that the character model does not reliably realise.

For the other languages, the rolling automatic results are reported as they exist. Bantawa has a complete single-system TTS report: 7,726 filtered utterances from 7,938 aligned verse segments, 22.6 hours after forced alignment, best checkpoint epoch 59 by MCD-DTW at 6.96 dB on six held-out verses, ECAPA speaker cosine 0.696, and Allosaurus phone-PER 0.591. Because no MMS Bantawa TTS model exists, these are checkpoint and pipeline diagnostics, not superiority claims. Magar has held-out automatic diagnostics at epoch 49 of a planned longer run: MCD 9.846 dB, PER 0.589, DNSMOS OVRL 2.937 versus 3.202 for natural reference, and speaker cosine 0.812. Its epoch-29 breathy-voice realisation probe is a null: median H1–H2 differs by only +0.06 dB, with a one-sided Mann–Whitney $p = 0.1295$ and sign test 8/17 cells. Gurung is also mid-training: its held-out MCD improved from 13.98 dB at warm start to 8.85 dB at epoch 10 and 8.87 dB at epoch 19. The latest epoch-19 onset probe finds /ɬ/ partly realised (cluster voiced fraction 0.70 versus 0.85 for plain /l/, $p = 0.042$), while the voiceless nasals remain null; the result is window-sensitive and therefore scoped as a diagnostic. Dhimal has alignment and Piper manifests ready—7,844 kept verse utterances at 99.35% alignment yield, plus 25,887 phrase-level clips and a 50-phone inventory—but no audited TTS checkpoint in the repository yet.

VI. Data, Release, and Limitations

The binding constraints are data provenance and speaker verification. The usable corpora are largely Bible, curriculum, dictionary, and newspaper material with restrictive or unclear redistribution rights, so the underlying text and audio are not released. We release frontends, evaluation harnesses, and derived tooling; model release depends on the source license of each language's audio. Native speakers remain the final authority on pronunciation. The current paper reports measured

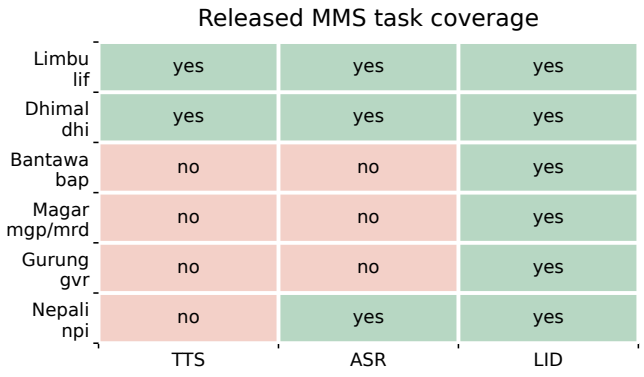


Fig. 4. MMS coverage relevant to the novelty claim. Magar, Bantawa, and Gurung have MMS LID coverage but no released MMS TTS or ASR checkpoint; Limbu and Dhimal do have MMS TTS/ASR.

results only: unrun speech contrast probes are not filled by analogy from another language.

VII. Related Work

This work sits between low-resource G2P, speech frontend engineering, and first-system speech work. Epitran and WikiPron provide broad multilingual pronunciation resources, but do not cover these Devanagari adaptations at the required scale [10], [11]. Brahmic schwa-deletion work for Hindi and Punjabi is the closest high-resource analogue, but our problem is the inverse: several target languages intentionally add or hide contrasts that Nepali does not have [13]. MMS demonstrates that massively multilingual TTS can cover many languages, including Limbu and Dhimal, but its released voices are character-level models rather than explicit, inspectable frontends [16]. For Magar, Bantawa, and Gurung, MMS provides language identification coverage but not released TTS or ASR models; this is why our novelty claim is first TTS/G2P for those languages, not first speech technology of any kind. We follow endangered-language and participatory NLP work in treating systems as community infrastructure whose claims must be scoped to the data and the evaluation that actually exist. We do not count broad zero-shot TTS catalog claims as language-specific prior art unless the ISO code and evaluated language match [19], [20], [21].

VIII. Conclusion

Shared script is not shared phonology. Across six languages of Nepal, a Nepali Devanagari frontend collapses the contrasts that the smaller languages need to preserve or recover. Explicit, inspectable G2P frontends make those decisions visible, testable, and reusable. The text-level contrast results are already strong; the speech results are intentionally more modest, with one robust Limbu length win, several honest nulls, and first-system diagnostics for languages with no released TTS comparator.

References

[1] P. Joshi *et al.*, "The state and fate of linguistic diversity and inclusion in the NLP world," in *ACL*, 2020.
[2] S. Bird, "Decolonising speech and language technology," in *COLING*, 2020.
[3] G. van Driem, *A Grammar of Limbu*. Berlin: Mouton de Gruyter, 1987.
[4] B. Michailovsky and M. Everson, "Revised proposal to encode the Limbu script in the UCS," ISO/IEC JTC1/SC2/WG2 N2410, 2002.
[5] M. Noonan, "Recent adaptations of the Devanagari script for the Tibeto-Burman languages of Nepal," in *Indic Scripts: Past and Future*, 2005.
[6] K. Grunow-Hårsta, *A Descriptive Grammar of Two Magar Dialects of Nepal*. Ph.D. dissertation, University of Wisconsin–Milwaukee, 2008.
[7] W. W. Glover and J. R. Glover, *A Guide to Gurung Tone*. Kathmandu: Tribhuvan University, 1972.
[8] M. Doornenbal, *A Grammar of Bantawa*. Ph.D. dissertation, Leiden University, 2009.
[9] J. T. King, *A Grammar of Dhimal*. Leiden: Brill, 2009.
[10] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for many languages," in *LREC*, 2018.
[11] J. L. Lee *et al.*, "WikiPron: Massively multilingual pronunciation mining," in *LREC*, 2020.
[12] J. Zhu, C. Zhang, and D. Jurgens, "ByT5 model for massively multilingual grapheme-to-phoneme conversion," in *Interspeech*, 2022.
[13] A. Arora, L. Gessler, and N. Schneider, "Supervised G2P conversion of orthographic schwas in Hindi and Punjabi," in *ACL*, 2020.
[14] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *ICML*, 2021.
[15] M. Hansen and Rhasspy contributors, "Piper: A fast, local neural text-to-speech system," 2023. https://github.com/rhasspy/piper
[16] V. Pratap *et al.*, "Scaling speech technology to 1,000+ languages," *JMLR*, vol. 25, 2024.
[17] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2021.
[18] X. Li, S. Dalmia, J. Li, M. Chou, A. W. Black, and F. Metze, "Universal phone recognition with a multilingual allophone system," in *ICASSP*, 2020.
[19] S. Zhang, B. Frey, and M. Bansal, "ChrEn: Cherokee-English machine translation for endangered language revitalization," in *EMNLP*, 2020.
[20] C. Taguchi *et al.*, "Killkan: The automatic speech recognition dataset for Kichwa," in *LREC-COLING*, 2024.
[21] W. Nekoto *et al.*, "Participatory research for low-resourced machine translation: A case study in African languages," in *Findings of EMNLP*, 2020.