When AI Plays the Cultural Game: AI-Mediated Regeneration of Penghu *Bokua* as Language Play in Digital Discourse

August Chao (趙逢毅)
Department of Computer Science and Information Engineering
National Penghu University of Science and Technology
Magong, Penghu, Taiwan
augchao@gms.npu.edu.tw

Abstract

Bokua (褒歌) is the improvisational call-and-response folk poetry of the Penghu Islands of Taiwan, performed in the endangered Penghu accent of Taiwanese Hokkien. As a constrained creative form—seven-character lines, tonal wordplay, and dialectal stance-taking—*bokua* represents a vivid instance of language play facing extinction as its speaker community rapidly declines. This paper presents **BokuaGen**, an AI system that mediates the regeneration of this endangered creative language in digital discourse. BokuaGen integrates bias-only speaker adaptation (0.13% of model parameters), kNN-based voice conversion, LLM-assisted forced alignment, and melody extraction from real *bokua* recordings to produce shareable Penghu-accent *bokua* videos.

Experiments on the only available Penghu-accent speaker (TIF1035, 234 utterances, NAER TAT-MOE corpus) show that (1) conventional fine-tuning collapses pitch (F0: 157→62 Hz), and adding an explicit F0 loss merely shifts the collapse to amplitude (RMS: 0.043→0.003)—a complementary dual mode collapse that destroys the dialect's creative tonal identity, while our bias-only approach recovers the target F0 (191 Hz); (2) age-controlled F0 comparison shows the Penghu speaker is indistinguishable from an age-matched Quanzhou speaker (both 191 Hz), consistent with their shared linguistic origin; (3) the complete pipeline achieves F1-band spectral alignment within 0.3 dB of the real speaker; (4) G-major pentatonic melody (G: 52%) extracted from real recordings provides culturally authentic accompaniment; (5) speaker-grounded G2P—aligning grapheme-to-phoneme conversion with the target speaker's actual readings rather than idealised dialect rules—raises syllable accuracy from 34.9% to 83.1%. The resulting digital artifacts—*bokua* speech synchronised with traditional melody—demonstrate how AI-mediated interaction can sustain endangered language play for contemporary audiences. A demo video is available at https://youtube.com/shorts/Mu0IAGK1ZbQ.

Keywords: language play, digital discourse, human-AI creativity, Penghu accent, *bokua*, endangered language preservation, text-to-speech, voice conversion

1 Introduction

Language play is central to human creativity and cultural identity (Cook, 2000). *Bokua* (褒歌), the traditional call-and-response folk poetry of the Penghu Islands of Taiwan, is a clear case of language play: performers improvise within strict formal constraints (seven-character lines, pentatonic melody, tonal rhyme) while serving social functions from courtship to community bonding (吳清地, 2019). Its creative power is inseparable from the Penghu accent of Taiwanese Hokkien, a conservative Quanzhou-lineage dialect (Ratte, 2009) whose tonal contours, entering-tone preservation, and prosodic rhythm constitute the aesthetic medium of the art form. We keep the two apart throughout: *bokua* is a verbal-art genre (verse form, melody, call-and-response practice), whereas the Penghu accent is the language variety in which it is performed. BokuaGen models them separately: the accent from Penghu-accent speech data, the genre from *bokua* verse constraints and melodies extracted from real *bokua* recordings.

Yet *bokua* is endangered. Taiwanese Hokkien is undergoing rapid intergenerational shift: in Taiwan's 2020 census, 65.9% of resi-

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 419–426
November 6-7, 2026 ©2026 Association for Computational Linguistics

dents aged 65+ but only 7.4% of those aged 6–14 reported it as their primary language (as reported in 立法院預算中心, 2023), and Ministry of Education statistics indicate a 60% decline in its use over three generations (Hioe, 2024). Penghu's population is markedly aged, with an aging index of 174.2 (residents aged 65+ per 100 aged 0–14; 澎湖縣政府主計處, 2021), which compounds this shift for its local accent. The scarcity is visible in the data: the NAER TAT-MOE corpus (國家教育研究院, 2023), the largest publicly available Taiwanese speech corpus (17 GB), contains only a *single* Penghu-accent speaker (TIF1035, 234 utterances). When that speaker community falls silent, so does *bokua*.

This paper asks: *Can AI mediate the regeneration of this endangered creative language in digital discourse?* We present **BokuaGen**, a modular TTS pipeline that generates Penghu-accent *bokua* and packages it as shareable MP4 artifacts, positioning AI as a creative partner in sustaining endangered language play for contemporary digital audiences. Our contributions span language technology, phonetics, and digital humanities.

2 Bokua as Language Play

Constrained creativity. *Bokua* operates within formal rules: four seven-character lines, tonal rhyme, and a call-and-response structure requiring the responding performer to improvise within the topic and metre set by the first. This mirrors the structure of other constrained creative traditions— rap battles, haiku, improv theatre—where constraint generates creativity (Boden, 2004).

Tonal wordplay. The Penghu accent preserves entering tones (-h/-p/-k/-t)—short, checked syllables ending in glottal stop or occlusive codas—at 18.2% of syllables (measured on the aligned TIF1035 corpus, Section 5.2). In this conservative Quanzhou-lineage variety (Ratte, 2009), these checked syllables contribute to the tonal puns and double meanings central to *bokua*'s wit. This phonological feature is not an incidental property of the dialect: it *is* the creative medium.

Dialectal stance-taking. Performing in Penghu accent marks the speaker as local, distinct from both Mandarin and standard Taiwanese (Bucholtz and Hall, 2005). AI regeneration in the same accent is thus not merely acoustic mimicry but a form of cultural stance re-enactment in digital discourse.

Digital discourse context. Traditionally performed face-to-face at fishing-community gatherings, *bokua* has migrated to YouTube and social media, losing its interactive call-and-response character in the process. BokuaGen restores this interactivity: users provide a topic and receive an AI-performed *bokua* response, recreating the creative dialogue in digital form.

3 Related Work

Taiwanese Hokkien speech synthesis. Pratap et al. (2023) provide the base model (`facebook/mms-tts-nan`, VITS architecture (Kim et al., 2021)) for 1,000+ languages including Taiwanese Hokkien. FormoSpeech (FormoSpeech Team, 2023) and the TAT corpus (國家教育研究院, 2023) provide speech data, but neither contains dialect-level annotation. BokuaGen is, to our knowledge, the first system targeting a specific Taiwanese sub-dialect (Penghu accent) for speech synthesis.

Low-resource speaker adaptation. AdaSpeech (Chen et al., 2021) adapts TTS to custom speakers by updating only layer normalisation parameters, conceptually related to our bias-only strategy. Jia et al. (2018) show that speaker embeddings learned from verification can be transferred to TTS. Our work extends this line by identifying and solving a previously unreported *dual mode collapse* in low-resource VITS fine-tuning.

Voice conversion and kNN-VC. kNN-VC (Baas et al., 2023) performs any-to-any voice conversion using WavLM-Large (Chen et al., 2022) features and HiFi-GAN (Kong et al., 2020), achieving strong speaker similarity without speaker-specific training. We apply kNN-VC as the second stage of our pipeline, using TIF1035 utterances as the target pool.

Endangered language and creativity in NLP. Bird (2020) argue for community-centred approaches to endangered language

technology. Studies of digital creativity in minority language contexts (Sebba, 2007) highlight how online communities adapt language play for new platforms. Recent work on AI-mediated creative writing (Gero et al., 2022) shows that AI can serve as a creative partner when users retain editorial agency over the final output. BokuaGen extends this framing to the domain of endangered oral creative traditions.

Taiwanese Hokkien phonetics and dialect analysis. Ratte (2009) is the only known quantitative study of Penghu Taiwanese phonology, documenting /g/-deletion rates in Magong City (60%+) and identifying a phonological conspiracy against voiced stops. We rely on his analysis for Layer 1 G2P corrections and to contextualise the F0 comparison in Section 6.3. The Formosa Speech Recognition Challenge (Liao et al., 2020) provided baseline ASR resources for Taiwanese Hokkien, and the TAT-MOE corpus (國家教育研究院, 2023) extended this with multi-speaker, multi-accent data. However, no prior synthesis work has targeted the Penghu accent specifically.

4 System Design Rationale

BokuaGen splits the low-resource task (234 utterances) into three stages, each targeting one bottleneck. *(1) Speaker-grounded G2P:* standard MOE romanisation diverges from TIF1035's actual readings (e.g., *Phînn-ôo* vs. *phênn-ôo*); grounding raises syllable accuracy from 34.9% to 83.1%. *(2) Bias-only adaptation:* full VITS fine-tuning collapses pitch (62 Hz) or amplitude; updating only bias terms (0.13% of parameters) recovers both. *(3) kNN-VC with post-processing:* training-free voice conversion moves timbre towards the speaker, bringing F1-band energy within 0.3 dB. Each stage follows established low-resource practice: parameter-efficient adaptation and training-free conversion (Chen et al., 2021; Baas et al., 2023).

5 The BokuaGen Pipeline

Data. TIF1035 is a 55-year-old female Penghu-accent speaker in TAT-MOE. Her 234 utterances (21.5 min, 16 kHz mono) are read speech: nine short utterances with numbers and dates, a twelve-turn scripted dialogue, and 213 phrase-level segments of three published Taiwanese-language prose essays, each with a Han-Lo transcript and a manual Tâi-lô romanisation. None is a *bokua* performance or verse; the corpus supplies the accent, not the genre.

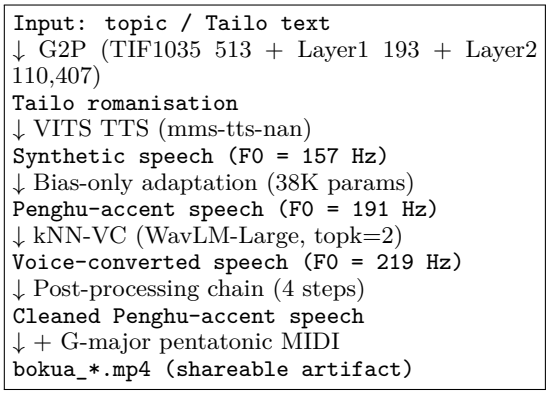

Figure 1: BokuaGen pipeline overview.

5.1 Speaker Adaptation: Solving Mode Collapse

Standard VITS fine-tuning on 234 utterances exhibits a *dual mode collapse* (Table 1): with mel-L1 loss alone, pitch collapses to 62 Hz (a sub-harmonic of the target 191 Hz); adding an explicit F0 loss fails to restore pitch (69 Hz) and instead collapses amplitude to near-silence (RMS 0.003). The root cause is a trivial solution to mel-L1: near-silent output minimises the average spectrogram distance across all training samples simultaneously. The two collapse modes are thus complementary—suppressing one surfaces the other—which is why we term the phenomenon dual.

We train only the flow bias parameters (38,248 / 30,006,560 = 0.13%) with amplitude regularisation:

$$\mathcal{L} = \mathcal{L}_{\text{mel}}(f_{\min} = 80 \text{ Hz}) + 2.0 \cdot |\log \hat{\sigma} - \log \sigma^*| \quad (1)$$

Training runs for 30 epochs with early stopping (patience 6); best mel loss 3.09; learning rate $5{\times}10^{-5}$.

Why bias parameters? In VITS, the flow module transforms the latent space via a series of affine coupling layers. The bias terms capture *additive* shifts in this latent representation and are analogous to the mean of a speaker's phonetic distribution. Intuitively, shifting bias values repositions the synthesis

Method	Params	F0	RMS
Original MMS	—	157 Hz	0.043
mel-L1 (30M)	30M	62 Hz	0.050
mel-L1 + F0 loss	30M	69 Hz	0.003
Bias-only (ours)	**38K**	**191 Hz**	**0.104**
TIF1035 (real)	—	191 Hz	—

Table 1: Ablation: fine-tuning strategy vs. F0 and amplitude. Only bias-only adaptation recovers the Penghu-accent tonal profile without mode collapse.

Metric	Value
Utterances aligned	234 / 234 (100%)
Total syllables	2,469
Mean confidence	0.699
Median confidence	0.760
Low confidence (<0.25)	174 (7.0%)
Mean syllable duration	123 ms
Syllables <40 ms	133 (5.4%)

Table 2: CTC forced alignment statistics for TIF1035 (234 utterances).

locus in acoustic feature space without distorting the manifold shape, which explains why 0.13% of parameters suffice to achieve a 34 Hz F0 shift. Weight matrices would need to restructure the feature manifold itself, creating far greater risk of instability with only 234 training samples.

5.2 CTC Forced Alignment

We align 234 utterances (2,469 syllables) using `torchaudio.functional.forced_align` with the MMS-1b-all nan CTC adapter. The key insight is that `forced_align` returns paths$[T_{\text{frames}}]$ where each element is a *vocabulary token ID*, forming *CTC runs* with strict 1:1 correspondence to target tokens, yielding reliable syllable-level timestamps (Table 2).

Low-confidence syllables are post-verified by a locally deployed LLM (GLM-4.7-flash via Ollama) with ±3-syllable context and Taiwanese phonological constraints, producing accept / adjust / flag decisions.

5.3 Voice Conversion and Post-Processing

kNN-VC uses 168 high-confidence utterances (conf $>$ 0.65, 45,991 WavLM frames) as the target pool. A four-step post-processing chain addresses specific artefacts (Table 3).

5.4 Speaker-Grounded G2P

Standard G2P dictionaries use the MOE romanisation (*Phînn-ôo* for 澎湖), which diverges from the target speaker (TIF1035: *phênn-ôo*). We build a layered G2P system and, crucially, ground it in the target speaker's actual pronunciation rather than idealised dialect rules.

A cautionary finding. We initially applied a systematic Quanzhou rule—the central-vowel shift (standard /i/ → /u/, the "heavy-u" feature)—drawn from classical rhyme books (彙音妙悟) and Wiktionary's Magong tags (e.g. 魚 *hî*→*hû*, 語 *gí*→*gú*). Evaluation against TIF1035's manual annotations revealed this rule is *over-applied*: in 72 cases it produced /u/ where TIF1035 actually reads /i/ (語 *gí*, 去 *khì*, 時 *sî*). The rhyme books encode an idealised, conservative Quanzhou variety, but the actual speaker—influenced by the dominant accent—retains /i/ for most characters. The "heavy-u" feature is *lexically conditioned* (specific 魚-rhyme words), not a blanket phonological rule.

Speaker-grounded layers. This motivates grounding G2P in the target speaker, because a TTS model can only synthesise sounds present in its training data: applying a rule the model never heard degrades synthesis quality. Our priority order is **(1)** TIF1035 readings (513 characters extracted from the 234 annotated utterances—the speaker ground truth); **(2)** Layer 1 Penghu-specific entries (193: place names, *bokua* vocabulary, Wiktionary white-readings such as *phênn-ôo*, *má-king*); **(3)** clean single-character standard readings (8,510 from 彙集雅俗通) and Layer 2 (110,407 ChhoeTaigi (ChhoeTaigi Contributors, 2020) plus three rhyme books)—with *no* blanket i→u rule.

Evaluation. We evaluate G2P intrinsically against TIF1035's manual romanisation (the

Artefact	Treatment	Result
HiFi-GAN comb	Spectral smooth (w=5)	cleaner harmonics
HF excess +4.7 dB	Shelf −4.7 dB@6 kHz	−93.6 → −96.7 dB
F1 bias +2.5 dB	FIR LTAS EQ (511-tap)	−47.8 → −50.6 dB
Residual noise	Spectral subtraction	−3 dB floor

Table 3: Post-processing chain. Target: TIF1035 F1 = −50.3 dB.

Metric	Blanket rules	Speaker-grounded
Syllable Acc.	34.9%	**83.1%**
Initial Acc.	74.2%	**96.7%**
Final Acc.	48.0%	**94.1%**
Tone Acc.	70.4%	**87.0%**
PER	35.1%	**6.4%**

Table 4: G2P intrinsic evaluation against TIF1035 manual annotation (1,487 syllables). Grounding G2P in the target speaker's actual readings substantially outperforms blanket Quanzhou rules, demonstrating that speaker-grounded G2P beats idealised dialect phonology for single-speaker TTS.

strongest available ground truth) at the syllable, initial, final, and tone levels (Table 4). Scoring needs a one-to-one syllable correspondence, so we evaluate on the 148 of 234 utterances whose G2P output has the same syllable count as the gold romanisation (1,487 syllables). Speaker-grounding raises syllable accuracy from 34.9% (blanket rules) to 83.1% and cuts the phoneme error rate from 35.1% to 6.4%. A train/test split (build the speaker table on half the utterances, test on the other half) yields 75.3% on seen characters and 46.2% on unseen characters, the latter bottlenecked by literary/colloquial reading ambiguity.

5.5 Melody Extraction from Real *Bokua*

Rather than algorithmically generating melody, we extract the tonal structure from a real *bokua* YouTube recording ("Penghu Silk and Bamboo—Passing on the Tradition", 2022). After pyin F0 extraction and quantisation to the nearest semitone, we find G-major pentatonic (G-A-B-D-E) as the dominant scale: G (52%), A (22%), B (15%), D (7%), E (4%). The near-complete pentatonic dominance (combined 100% of long notes) confirms that *bokua* uses pentatonic melody exclusively, in contrast with the algorithmically generated MIDI used in our earlier prototype, which contained chromatic passing tones.

A rule-based four-line melody is constructed from the extracted scale with ascending phrase for line 1 (question), stable middle register for line 2 (answer), upper-register leap for line 3 (re-question), and descending cadence to G for line 4 (resolution). This mirrors the traditional *bokua* call-and-response arc.

6 Experiments and Results

6.1 System Comparison

Table 5 summarises objective metrics across five system configurations, evaluated on Sentence 1 ("媽宮古城上蓋鬧", 4.8 s) from the benchmark set. F0 is the mean pyin estimate over voiced frames; RMS is the root-mean-square amplitude of the waveform (full scale = 1); F1 and HF are the mean Welch power spectral density in dB over the first-formant band (200–1000 Hz) and the high-frequency band (6–8 kHz). On all four, closer to the TIF1035 row is better.

Key observations: (1) Mode collapse is evident in mel-L1 fine-tune (F0 = 62 Hz), destroying the Penghu-accent tonal contour. (2) Bias-only adaptation recovers F0 (193 Hz ≈ the speaker's corpus-mean target 191 Hz) at the cost of slightly higher RMS. (3) kNN-VC brings spectral characteristics closer to TIF1035 (pre-EQ F1: −47.8 vs. −50.3 dB; HF: −93.6 vs. −95.1 dB). (4) Post-processing further reduces the F1 gap to 0.3 dB (−50.6 vs. −50.3 dB, Table 3).

6.2 Bokua Generation Output

Table 6 reports the final BokuaGen output for two *bokua* songs.

Both songs maintain F0 in the 200–215 Hz range, consistent with the Penghu-accent female speaker profile and within 12% of TIF1035's corpus-mean F0 (191 Hz; per-song deviations 4.7–11.5%). The mixed versions overlay the voice signal with G-major pen-

System	F0	RMS	F1	HF
Original MMS	157 Hz	0.136	−47.6	−88.2
mel-L1 fine-tune†	62 Hz	0.156	−46.4	−76.2
Bias-only (ours)	193 Hz	0.235	−42.8	−90.4
+kNN-VC +post-proc	219 Hz	0.129	−50.6	−96.7
TIF1035 (real)	163 Hz	0.105	−50.3	−95.1

Table 5: System comparison (F0, RMS, F1/HF spectral energy in dB), evaluated on Sentence 1; the TIF1035 row is this single utterance (F0 163 Hz), whereas the speaker's corpus-mean F0 is 191 Hz (Table 7). † mel-L1 fine-tune (30M params, cf. Table 1) exhibits pitch collapse despite reasonable RMS. kNN-VC + post-processing best approximates TIF1035 spectrum.

Song	F0	RMS	Duration
Penghu Haijing (voice)	205 Hz	0.172	8.5 s
Penghu Haijing (mixed)	200 Hz	0.172	9.0 s
Yutsun Shenghuo (voice)	213 Hz	0.157	10.9 s
Yutsun Shenghuo (mixed)	213 Hz	0.145	11.4 s

Table 6: BokuaGen final output: two *bokua* songs in Penghu accent with G-major pentatonic melody overlay.

tatonic melody (BPM=72) extracted from a real YouTube *bokua* performance, with voice/melody ratio 85:15. The final MP4 output (h264 + AAC, 9.0–11.4 s) is designed for direct sharing on social media platforms.

6.3 F0 Dialect Analysis

Comparing TIF1035 against the corpus's eight Quanzhou-accent speakers (five representative speakers shown in Table 7), we find that the apparent F0 difference is consistent with an age effect (younger women have higher F0). After age-matching, TIF1035 coincides with IUF1012 (53F; both 191 Hz), while the other 53-year-old female, IUF1020, is 26 Hz higher (217 Hz), so individual variation within the same age–sex group is not negligible. The comparison is thus consistent with, though not proof of, Ratte (2009)'s characterisation of Penghu dialect as a conservative Quanzhou-lineage variety. This finding supports using Quanzhou-accent data as a training supplement for Penghu-accent TTS.

6.4 Error Analysis

Table 8 shows the most common failure modes in the alignment and synthesis pipeline.

The entering-tone duration errors are linguistically expected: entering tones (-h/-p/-k/-t) are short by phonological definition, and pyin's minimum viable window (40 ms) systematically underestimates their boundaries. LLM post-verification accepts most of these as correct with the rule *"entering tones below 80 ms are acceptable."*

7 Discussion: AI as Creative Interlocutor

BokuaGen repositions AI from a text-processing tool to a *creative interlocutor* in an endangered language play tradition. Three implications for the study of creativity and digital discourse:

Constraint as creative scaffold. The formal constraints of *bokua*—pentatonic melody, seven-character lines, tonal rhyme—function as a scaffold that channels AI creativity, analogous to constrained decoding in NLP. Formalised creative traditions may be more amenable to AI regeneration than free-form discourse, because the constraints define the creative space.

Dialectal features as creative medium. The Penghu accent's tonal profile is not incidental to *bokua*: it *is* the creative medium. Our finding that 0.13% of parameters suffices to capture the speaker's F0 profile suggests that speaker identity may be encoded in surprisingly few parameters, with implications for other endangered dialect creative traditions.

Limitations and future work. The current system lacks subjective (MOS) evaluation by native Penghu speakers, which remains essential for validating perceptual authenticity.

Speaker	Accent	Age/Sex	F0
TIF1035	Penghu	55F	191 Hz
IUF1012	Quanzhou	53F	**191 Hz**
IUF1020	Quanzhou	53F	217 Hz
IUF1026	Quanzhou	23F	248 Hz
IUM1017	Quanzhou	60M	168 Hz
IUM1007	Quanzhou	41M	126 Hz

Table 7: Age-controlled F0 comparison (pyin, bandpass 75–500 Hz). TIF1035 coincides with the age-matched Quanzhou speaker IUF1012 (both 191 Hz); the other 53-year-old female, IUF1020, is 26 Hz higher (single-speaker comparison; the corpus has only one Penghu speaker).

Error type	Rate	Example / Cause
Short syllables (<40 ms)	5.4%	Entering tones (-h/-p/-k); pyin underestimates duration
Low CTC conf (<0.25)	7.0%	OOV phonemes; /ts/ clusters not in MMS training data
F0 overshoot (+28 Hz)	systematic	kNN-VC output 219 Hz vs. speaker mean 191 Hz; not explained by the target pool, whose mean F0 matches the speaker's
HF excess (+4.7 dB)	systematic	HiFi-GAN vocoder artefact; removed by shelf filter

Table 8: Error analysis: dominant failure modes and remedies.

We plan to conduct a listening study with speakers from Penghu County as evaluators, using the five-point MOS scale with specific attention to *bokua* tonal authenticity and dialectal naturalness.

The single-speaker training data (234 utterances, 21.5 min) limits prosodic diversity. Incorporating the full Quanzhou-accent subset (2,323 utterances, 8 speakers) as a pre-training stage, followed by bias-only adaptation on TIF1035, could improve coverage of prosodic variation within the dialect. Additionally, a GAN discriminator trained on Penghu-accent samples could potentially allow fine-tuning more parameters (entire flow module, ∼2M) without mode collapse, closing the remaining 0.3 dB F1 gap.

On the creative side, future work will explore interactive *bokua* generation: users provide the opening two lines ("call") and the AI generates a contextually appropriate response ("answer"), realising the full call-and-response structure in digital discourse. A demo of the current system is available at `https://youtube.com/shorts/MuOIAGK1ZbQ`.

7.1 Implications for Language Play Research

BokuaGen contributes to the broader study of language play and creativity in digital discourse in three ways.

First, it provides a *computational model of constrained creativity*: the formal constraints of *bokua* (tone, metre, rhyme) can be operationalised as hard constraints in text generation and as target distributions in speech synthesis. This suggests a methodology for studying other constrained creative traditions computationally—limerick, haiku, call-and-response spiritual— by treating the constraints as generative structure rather than limitation.

Second, it raises questions about *authenticity in AI-mediated creativity*. When is an AI-generated *bokua* "authentic"? Our spectral analysis quantifies acoustic proximity to a real speaker, but authenticity in language play also involves pragmatic appropriateness, cultural knowledge, and social recognition by the community. Future work could use the BokuaGen output as a stimulus in listener judgement studies asking: does this sound like *bokua*? Does it sound like Penghu?

Third, the system provides a *digital discourse artefact*: the MP4 video is not just a playback file but a culturally positioned statement— an AI performing in a minority dialect's creative register for a global digital audience. This raises productive questions for discourse analysis: Who is the "speaker"? What stance does the AI perform? How do viewers interpret the interaction between authentic visual imagery (the watercolour illustration) and AI-generated voice in the target dialect?

8 Conclusion

We have presented BokuaGen, demonstrating that AI can mediate the regeneration of *bokua* as an endangered form of language play in digital discourse. The key contribution—bias-only speaker adaptation with amplitude regularisation—solves the previously unreported dual mode collapse in low-resource VITS fine-tuning, preserving the dialectal tonal features essential to *bokua*'s creative identity. The age-controlled F0 analysis, CTC-runs alignment method, and diagnosis-driven post-processing chain are all independently reusable components for related work. More broadly, BokuaGen provides a framework for AI-mediated preservation of constrained creative traditions in endangered dialects, at the intersection of digital humanities, computational creativity, and minority language technology.

References

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. kNN-VC: A training-free, any-to-any voice conversion via nearest neighbors retrieval from off-the-shelf discrete speech representations. In *Proceedings of Interspeech*.

Steven Bird. 2020. Decolonising speech and language technology.

Margaret A. Boden. 2004. *The Creative Mind: Myths and Mechanisms*, 2nd edition. Routledge.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4–5):585–614.

Mingjian Chen, Xu Tan, Bohan Li, et al. 2021. AdaSpeech: Adaptive text to speech for custom voice. In *Proceedings of ICLR*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

ChhoeTaigi Contributors. 2020. ChhoeTaigi Taiwanese dictionary database. https://github.com/ChhoeTaigi/ChhoeTaigiDatabase.

Guy Cook. 2000. *Language Play, Language Learning*. Oxford University Press.

FormoSpeech Team. 2023. FormoSpeech: Taiwanese speech resources. https://huggingface.co/formospeech.

Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*.

Brian Hioe. 2024. Polling shows language remains contested in Taiwan. New Bloom Magazine, 6 April 2024. https://newbloommag.net/2024/04/06/polling-native-languages/.

Ye Jia, Yu Zhang, Ron J. Weiss, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *arXiv preprint arXiv:2106.06103*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*.

Liang-Hsuan Liao et al. 2020. Formosa speech recognition challenge 2020 and Taiwanese across Taiwan corpus. In *Proceedings of Interspeech*.

Vineel Pratap, Andros Tjandra, Bowen Shi, et al. 2023. Scaling speech technology to 1,000+ languages. In *Proceedings of ACL*.

Alexander T. Ratte. 2009. A dialectal and phonological analysis of Penghu Taiwanese. B.A. Honors Thesis, Williams College.

Mark Sebba. 2007. Spelling and society: The culture and politics of orthography around the world.

吳清地. 2019. 澎湖褒歌的文化意涵與傳承. 澎湖文獻資料彙編.

國家教育研究院. 2023. 臺灣台語語料庫應用檢索系統 (TAT-MOE). https://tggl.naer.edu.tw/. Accessed: 2026.

澎湖縣政府主計處. 2021. 澎湖縣統計月報（110年10月）. https://www.penghu.gov.tw/ch/home.jsp?id=10178.

立法院預算中心. 2023. 113年度中央政府總預算案整體評估報告. https://www.ly.gov.tw/Pages/Detail.aspx?nodeid=46405&pid=231979.