LEARN WHAT TO SAY: GENERALIZABLE ACTIVE ENROLLMENT FOR FEW-SHOT VOICE PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot voice personalization can reproduce a target speaker from only a few seconds of reference audio, yet users receive little guidance about what they should say. We formulate *active voice enrollment* as a budget-constrained sequential selection problem and introduce **LEAP** (**L**anguage-Transferable **E**nrollment via **A**ctive **P**olicy Learning). LEAP predicts the marginal utility of each candidate utterance conditioned on the content already selected, enabling it to construct complementary enrollment sets without synthesizing speech or measuring downstream quality at inference time. We evaluate LEAP across Korean, Japanese, and Vietnamese, multiple voice-personalization models, matched recording budgets, and human listening studies. LEAP improves automatic personalization utility over random selection, natural user-selected enrollment, phoneme coverage, language-specific handcrafted objectives, and a learned static-ranking baseline. Crucially, when trained only on outcomes from two source languages, LEAP outperforms a handcrafted objective designed specifically for the held-out target language; its gains also persist for an unseen voice-personalization model. Human listeners prefer outputs generated from actively selected enrollment for speaker similarity, naturalness, and content clarity. These results show that enrollment utility contains transferable structure: when only a few seconds are available, performance depends not only on how much a user speaks, but also on what the system asks them to say. The selection summary and lexicon file for LEAP are available at `https://anonymous.4open.science/r/Learn-What-to-Say-F86C`.

1 INTRODUCTION

Modern speech foundation models can personalize a voice from only a few seconds of reference audio, enabling personalized speech generation (Jia et al., 2018; Casanova et al., 2022), accessibility, dubbing (Cong et al., 2024), and human–computer interaction. Progress has largely focused on the model: speaker representations, generative architectures, multilingual pretraining, and adaptation objectives (Jia et al., 2018; Ju et al., 2024; Chen et al., 2025; Casanova et al., 2022; Hussain et al., 2025). The reference speech itself is usually treated as a fixed input. Users are asked to provide a short recording, but receive little guidance about what they should say. This omission matters because equal-duration recordings are not equally informative. A five-second recording may repeatedly realize similar sounds, or it may provide diverse phonetic contexts, pronunciation processes, prosodic contrasts, and transitions that are useful for personalization. Moreover, an utterance's value depends on what has already been recorded. A sentence that is informative in isolation may add little after similar evidence is present, while an otherwise ordinary sentence may supply precisely the evidence missing from the current enrollment. Thus, voice enrollment is not only a question of how much speech is available, but a sequential decision problem: under a limited recording budget, what should the system ask the user to say next? Existing reference-selection methods mainly retrieve recordings from an available speech database or optimize acoustic and stylistic compatibility with a target generation task (Xue et al., 2024; Jiang et al., 2024). Classical text-selection methods construct phonetically balanced scripts for speech-corpus development (van Santen & Buchsbaum, 1997; Bozkurt et al., 2003; Nose et al., 2015; Kwon et al., 2025). Neither setting directly addresses few-shot voice personalization. At a budget of only several seconds, corpus-scale coverage is unattainable; nor can phoneme coverage alone determine which evidence complements a partially recorded enrollment set or best supports a modern personalization model. Handcrafted linguistic objectives provide a useful

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

starting point, but assign value according to predefined inventories. They do not learn conditional interactions among selected utterances, their downstream utility, or whether useful selection principles transfer across languages and voice models. We formulate this problem as *active voice enrollment*: selecting an ordered set of utterances that maximizes downstream personalization quality subject to a recording-time budget. At each step, an enrollment policy observes the linguistic evidence already selected, the remaining candidate utterances, and the available budget. It predicts each candidate's conditional marginal utility and selects the candidate with the highest predicted additional benefit per second. Unlike conventional reference retrieval (Xue et al., 2024; Jiang et al., 2024), the selected utterances need not exist as recordings in advance: the system determines what a new user should say before enrollment audio is collected. The central challenge is generalization. A selector trained and evaluated in one language with one voice model may exploit language-specific lexical regularities or the idiosyncratic preferences of that architecture. Its practical value would then be limited, requiring new downstream experiments whenever a language or personalization model changes. We instead ask whether the linguistic evidence useful for enrollment has transferable structure. **Can a policy trained on some languages identify useful utterances in a language for which it has never observed personalization outcomes? Can it remain effective when the downstream voice model changes? And can it outperform a handcrafted linguistic objective designed specifically for the unseen language?**

To answer these questions, we introduce a multilingual active-enrollment benchmark spanning Korean, Japanese, and Vietnamese. These languages exhibit complementary phonological systems, including distinct syllable structures, consonantal constraints, context-dependent pronunciation processes, and tonal or pitch-based contrasts. For each language, we construct a linguistically curated candidate lexicon annotated with phonemic, syllabic, morphophonological, prosodic, and transition-level information. The benchmark uses matched recording budgets, speaker-disjoint partitions, multiple few-shot voice-personalization models, automatic evaluation, and human perceptual judgments. It supports controlled comparisons among random selection, natural user-selected speech, phoneme coverage, language-specific handcrafted utility, learned selection, and an empirical oracle. We propose LEAP (**L**anguage-Transferable **E**nrollment via **A**ctive **P**olicy Learning), a budget-aware, set-conditioned policy for active voice enrollment. Rather than assigning each candidate a fixed informativeness score, LEAP predicts its marginal utility relative to the current enrollment set. It represents utterances through language-comparable phonological, prosodic, and transition-level evidence, and explicitly distinguishes evidence that is already present from evidence that a candidate would newly contribute. This lets the policy model complementarity: a candidate is valuable when it provides evidence not already supplied by the selected content. LEAP is trained offline using relative utility supervision derived from downstream personalization outcomes across speakers, languages, and voice models. At selection time, however, it uses only the candidate inventory and text-side linguistic annotations; it requires no synthesized outputs, downstream quality measurements, or personalization runs.

We evaluate LEAP under transfer to unseen speakers, recording budgets, languages, and voice-personalization models. Across these settings, LEAP outperforms random enrollment, natural user-selected speech, conventional phoneme coverage, language-specific handcrafted utility, and learned static ranking. In particular, when transferred to a held-out language, LEAP surpasses a handcrafted linguistic objective constructed using annotations from that language, despite never observing its personalization outcomes during training. Its advantage also persists on previously unseen voice-personalization models. Human evaluations show that these gains are perceptually meaningful under matched recording budgets, including speaker similarity, naturalness, and content clarity. Analyses further show that the improvement reflects conditional complementarity beyond utterance length, rare-phoneme preference, or standalone utility.

Our contributions are four-fold: 1) We formulate few-shot voice enrollment as a budget-constrained active selection problem, shifting the focus from how much reference speech is available to what a user should say; 2) We introduce LEAP, a set-conditioned active policy that learns conditional marginal utility from complementary phonological, prosodic, and transition-level evidence, without running the downstream voice model at selection time; 3) We present a multilingual active-enrollment benchmark spanning Korean, Japanese, and Vietnamese, with multiple voice-personalization models, matched recording budgets, speaker-disjoint evaluation, and human perceptual judgments; 4) We demonstrate transfer to unseen recording budgets, lexical realizations, voice-personalization models, and languages, including outcome-free transfer to a held-out language where LEAP exceeds a target-language handcrafted objective.

2 Related Work

Few-Shot Voice Personalization Few-shot and zero-shot voice-personalization systems reproduce an unseen speaker from a short reference recording. Early work transferred speaker representations from speaker verification to multi-speaker synthesis (Jia et al., 2018), and subsequent systems demonstrated zero-shot speaker adaptation and voice conversion across speakers and languages (Casanova et al., 2022). Recent approaches improve reference-conditioned generation through latent diffusion, neural codec language modeling, flow matching, disentangled modeling of speaker identity and prosody, or adaptation using downstream quality or preference signals (Shen et al., 2024; Ju et al., 2024; Zheng et al., 2025; Chen et al., 2025; Lee et al., 2025; Zhang et al., 2025; Li et al., 2025; Hussain et al., 2025). These methods improve how a synthesis model represents or adapts to an *available* reference recording. We address the complementary problem that arises before that recording exists: under a short recording budget, what should a new user say to provide the most useful evidence for personalization?

Multilingual Speech Synthesis and Cross-Lingual Transfer Multilingual voice-cloning systems aim to preserve speaker identity while generating speech across languages with different phonological and prosodic systems (Casanova et al., 2022; Le et al., 2023; Casanova et al., 2024; Zheng et al., 2025; Hu et al., 2026). This line of work studies cross-lingual generalization of the *synthesis model*. In contrast, LEAP studies whether an *enrollment policy* trained from personalization outcomes in some languages can identify informative recording content in an unseen language without observing its outcomes, thereby learning transferable enrollment utility rather than applying a multilingual synthesizer to new text.

Text, Reference, and Active Selection Classical speech-corpus construction selects sentences to maximize phonetic coverage, balance contextual units, or reduce redundancy (van Santen & Buchsbaum, 1997; Bozkurt et al., 2003; Nose et al., 2015; Kwon et al., 2025). These methods target population-level corpus coverage across many recordings, rather than the few seconds available for personalizing one speaker. Text-dependent speaker verification and pass-phrase verification likewise study enrollment utterances, but typically verify whether a user produced a prescribed phrase or combine already recorded enrollment segments rather than choosing a personalized recording script (Zeinali et al., 2018; Ai et al., 2026). A separate line retrieves or aggregates *existing* recordings to improve their acoustic, stylistic, or contextual compatibility with a target synthesis request (Xue et al., 2024; Jiang et al., 2024). More generally, active-learning and budgeted subset-selection methods select already available examples under a limited acquisition or inclusion cost (Sener & Savarese, 2018; Mirzasoleiman et al., 2020). Active voice enrollment differs from these settings. It selects *future user utterances* before candidate audio exists, so selection cannot use candidate acoustics, synthesized outputs, or downstream quality at inference time. Further, an utterance's value is conditional on the current enrollment set: a high-utility utterance in isolation may be redundant, while another may supply missing phonological, prosodic, or transition-level evidence. LEAP therefore learns a set-conditioned marginal-utility function from offline personalization outcomes, but deploys using only text-side linguistic annotations. We evaluate whether this learned selection rule transfers across speakers, languages, recording budgets, and voice-personalization models.

3 Method

We formulate *active voice enrollment* as selecting what a new user should record before any enrollment audio exists. Given a language-specific pool of candidate utterances, LEAP selects a complementary set under a recording budget using text-side linguistic evidence alone at selection time. Its central design choice is to score an utterance by its *marginal* value relative to evidence already selected, rather than by a fixed utterance-level ranking.

3.1 Active Voice Enrollment

Let $\mathcal{C}^\ell = \{x_1, \ldots, x_N\}$ be an unrecorded candidate inventory for language ℓ. Each candidate x_i has an estimated recording duration $\hat{d}_i$ and a structured linguistic annotation $A(x_i)$. Under a recording-time budget B, we construct $S \subseteq \mathcal{C}^\ell$ such that

$$\sum_{x_i \in S} \hat{d}_i \leq B.$$

For target speaker s and downstream voice-personalization model m, let $Q(S; s, m)$ denote the quality obtained by personalizing m with recordings of S. The value of a candidate is conditional on

the current enrollment set S_t:

$$\Delta Q(x_i \mid S_t; s, m) = Q(S_t \cup \{x_i\}; s, m) - Q(S_t; s, m).$$

Computing this quantity during enrollment would require recording every candidate and repeatedly running personalization, synthesis, and evaluation. LEAP instead learns a predictor of conditional marginal utility,

$$\hat{u}_{i,t} = \widehat{\Delta Q}_\theta(x_i \mid S_t, b_t), \qquad b_t = B - \sum_{x_j \in S_t} \hat{d}_j,$$

where b_t is the remaining budget. The predictor receives only the linguistic evidence in x_i and S_t, together with budget information; it receives neither candidate audio nor target-speaker, downstream-quality, synthesized-output, or voice-model identity at selection time.

3.2 Set-Conditioned Evidence Policy

Each candidate inventory contains text, pronunciation information, estimated duration, and annotations for segmental, structural, prosodic, pronunciation-process, transition, and articulatory–visemic evidence. We map language-specific annotations to shared functional families $\mathcal{G}$, while retaining a language-specific applicability mask for families unavailable in a language. The detailed inventories and annotation mappings are provided in Appendix A.1.

For an event $a \in A_g(x_i)$ from family g, a deterministic canonical descriptor $v_g(a)$ represents its functional role rather than its surface lexical identity, and a family-specific encoder produces $z_g(a) = E_g(v_g(a))$. For an event collection T, let $e_g(T) = \mathrm{Pool}\{z_g(a) : a \in T\}$, where Pool denotes permutation-invariant mean–max pooling. At step t, we form novel and overlapping event collections by canonical-descriptor set difference and intersection between $A_g(x_i)$ and $\bigcup_{x \in S_t} A_g(x)$. Thus, for every family, the policy represents the candidate evidence, the evidence already present in S_t, and the candidate evidence that is respectively new or overlapping with S_t:

$$e_g(x_i), \quad e_g(S_t), \quad e_g^{\mathrm{new}}(x_i \mid S_t), \quad e_g^{\mathrm{overlap}}(x_i \mid S_t).$$

The policy also receives normalized coverage, novelty, and redundancy statistics for each family. Hence, it can distinguish both *which* linguistic events are complementary and *how much* additional evidence a candidate contributes. The complete state–candidate representation is

$$h_{i,t} = \left[\left\{ \phi_g(x_i, S_t, \rho_g^\ell) \right\}_{g \in \mathcal{G}} ; \mathbf{q}(x_i); \frac{\hat{d}_i}{B}; \frac{b_t}{B}; \frac{|S_t|}{|\mathcal{C}^\ell|} \right], \qquad \hat{u}_{i,t} = f_\theta(h_{i,t}),$$

where ϕ_g concatenates the four pooled embeddings, three scalar statistics, and the applicability mask for family g; $\mathbf{q}(x_i)$ contains normalized count features; and f_θ is a shared multilayer scorer. There is no explicit language, speaker, or downstream-model identifier. The applicability masks retain only the distinctions needed to interpret the available annotations. Further details about the policy can be found in Appendix A.2.

3.3 Offline Utility Learning and Outcome-Free Transfer

LEAP is trained offline from source-language speakers and source voice-personalization models. For each sampled state–candidate pair, its target is the empirical marginal utility above, where

$$Q(S; s, m) = \sum_{k=1}^{K} \alpha_k \widetilde{M}_k(S; s, m).$$

The fixed measures $\widetilde{M}_k$ capture speaker similarity, content preservation, and predicted naturalness; they are oriented so that higher is better and standardized using training speakers only. We normalize marginal utilities within each source language–model condition and train f_θ with a listwise ranking objective over candidates evaluated in the same state, together with Huber regression on normalized utility. This teaches both which candidate is preferable given the current set and the expected magnitude of its contribution. The sampled trajectories, target construction, normalization, and full objective are specified in Appendix A.3. At deployment, candidates are still unrecorded. From the feasible set

$$\mathcal{F}_t = \left\{ x_i \in \mathcal{C}^\ell \setminus S_t : \hat{d}_i \le b_t \right\},$$

LEAP greedily selects

$$x_t^* = \arg\max_{x_i \in \mathcal{F}_t} \frac{\hat{u}_{i,t}}{\hat{d}_i^\gamma}, \qquad S_{t+1} = S_t \cup \{x_t^*\}.$$

Here $\gamma \geq 0$ controls the duration preference and is selected using source-language validation only. Selection stops when no candidate fits the remaining budget. Constructing an enrollment script therefore requires only candidate-annotation preprocessing and lightweight policy evaluation; it never runs a personalization model, synthesizes speech, or measures downstream quality. For outcome-free transfer to an unseen language, the language supplies an eligible candidate inventory and its text-side linguistic annotations only. The same canonical descriptors, applicability masks, and set-conditioned features are constructed without downstream personalization outcomes, checkpoint selection, or policy updates. A held-out voice-personalization model likewise supplies neither utility outcomes nor a model identifier.

4 Experimental Setup

We evaluate whether active enrollment improves few-shot voice personalization under matched recording budgets and whether its selection principles generalize across speakers, budgets, languages, and downstream voice-personalization models. All comparisons use speaker-disjoint partitions, identical candidate pools, estimated-duration constraints, reference construction, source recordings, downstream inference settings, and evaluation items.

4.1 Benchmark and Recordings

Our primary benchmark comprises Korean, Japanese, and Vietnamese, which expose complementary phonological structure: context-dependent consonantal processes and coda behavior in Korean, moraic timing and pitch accent in Japanese, and lexical tone with language-specific syllable constraints in Vietnamese. Each language has a lexicon of 96 candidate prompts, yielding 288 candidates in total. Each entry includes its text, pronunciation information, estimated duration, and structured linguistic annotations. Appendix B.1 details lexicon construction and the language-specific annotation mappings. We collect full-lexicon recordings from 30 native speakers per language (90 speakers total) and use mutually exclusive 18/6/6 train/validation/test splits within each language. Offline utility supervision is collected only from the 54 training speakers. Validation speakers are used solely for early stopping and checkpoint selection; test-speaker recordings and personalization outcomes are not used during policy development. Recording, preprocessing, and quality-control procedures are described in Appendix B.1.1.

4.2 Voice-Personalization Protocol and Budgets

We evaluate three frozen zero-shot voice-personalization systems with distinct content and speaker-representation mechanisms: Seed-VC, OpenVoice V2, and FreeVC. For a test speaker, each system receives a selected target-speaker reference set and a disjoint source recording containing the desired linguistic content. Only the reference set varies across selection methods. We generate 20 converted utterances per test speaker; these utterances are disjoint from the enrollment lexicon. Exact checkpoints, preprocessing, and inference configurations are deferred to Appendix B.2. We evaluate 3-, 5-, 7-, and 10-second estimated recording budgets. All methods use the same feasibility constraint based only on the pre-recording estimate $\hat{d}_i$; speaker-specific realized durations are unavailable to the selector. We report realized duration, overrun rate, and deviation from the nominal budget as protocol diagnostics. Appendix B.2.1 gives the full budget-handling rule.

4.3 Baselines and Supervision

We compare LEAP with uniform random selection, natural user-selected enrollment, phoneme coverage, language-specific handcrafted utility, and a learned static ranker. The static ranker retains the same text-side inputs as LEAP but removes current-set, novelty, overlap, and remaining-budget representations. We additionally report a non-deployable empirical greedy oracle solely to quantify headroom. Baseline definitions and tie-breaking rules are given in Appendix B.3. Policy targets use an equal-weight combination of standardized speaker similarity, content preservation, and predicted naturalness. We collect bounded state–action comparisons from random, coverage-based, and exploratory trajectories. Target normalization, action collection, architecture, and optimization details are in Appendix B.3.1.

4.4 Generalization Protocols

All primary results use held-out test speakers. For cross-budget transfer, we perform leave-one-budget-out training over the four budgets, with no fine-tuning or budget-specific checkpoint selection at

Table 1: **Language-specific components of the handcrafted enrollment baseline.** Each baseline combines the applicable components with uniform weights; LEAP instead uses shared text-side linguistic evidence. Marked struct. denotes marked structure, Morph denotes morphophonology, and tr. denotes transitions, respectively.

Component	Korean	Japanese	Vietnamese	English
Segmental	Phonemes	Phonemes	Phonemes	Phonemes
Marked struct.	Codas	Special morae	Final consonants	Codas and consonant clusters
Contrast	Morph	Pitch accent	Lexical tone	Lexical stress and vowel reduction
Sequential	Phoneme tr.	Phoneme/mora tr.	Phoneme/tone tr.	Phoneme tr.

the excluded budget. For leave-one-language-out transfer, we train on outcomes from two primary languages and apply the policy to the third language's candidate lexicon and text-side annotations. No target-language recordings, outcomes, normalization statistics, validation scores, or outcome-derived rankings are used during policy development. We additionally transfer a policy trained on all three primary languages to English under the same outcome-free condition. For leave-one-model-out transfer, we train using utility labels from two of the three voice-personalization models and evaluate selection with the excluded model. All normalization, early stopping, and checkpoint selection use only source models. Full fold definitions are listed in Appendix B.4.

4.5 Evaluation and Statistical Analysis

For each test speaker, model, budget, and enrollment method, we measure target-speaker similarity, content preservation, and predicted naturalness using frozen ECAPA–TDNN embeddings, Whisper large-v3 transcriptions, and NISQA–TTS, respectively. Scores are averaged within speaker and macro-averaged across speakers; we report language-specific results and unweighted macro-averages. Metric implementations are detailed in Appendix B.4.1. Our confirmatory comparison pairs LEAP with the pre-specified strongest deployable non-oracle baseline in each setting. We use paired speaker-level bootstrap tests with 10,000 resamples, with hierarchical resampling for comparisons involving random selection. Human evaluation at 3 and 7 seconds uses blinded pairwise judgments of speaker identity, naturalness, content clarity; preferences are analyzed with a mixed-effects logistic regression. We report 95% confidence intervals, effect sizes, and Holm–Bonferroni-corrected p-values. Full human-evaluation, exclusion, and ethics protocols appear in Appendix B.4.2.

4.6 Leakage Control

Test-speaker recordings and outcomes are excluded from policy training, normalization, early stopping, and hyperparameter selection. In leave-one- language-out and leave-one-model-out experiments, all outcomes and validation signals from the held-out language or model are likewise excluded throughout development. At selection time, policies use text-side annotations and estimated durations only; all deployable methods share the same candidate pools, budgets, source recordings, reference construction, inference settings, and evaluation items.

5 Results

This section reports the primary evaluations. Additional evaluations of generalization to unseen speakers, recording budgets, and lexical realizations; transfer across voice-personalization models; and the linguistic evidence selected by LEAP are provided in the Appendix C.

5.1 Active Enrollment Improves Personalization Under Recording Budgets

We first ask whether selecting *what* a user records improves few-shot voice personalization under matched estimated recording-time budgets. Table 2 reports normalized composite personalization utility, averaged across Korean, Japanese, and Vietnamese, for four maximum recording budgets. All methods use the same speaker-disjoint test speakers, downstream voice-personalization models, and candidate inventory; they differ only in the enrollment utterances selected under the available budget. LEAP achieves the highest utility among outcome-free deployable methods at every budget. Its advantage is largest under the most restrictive enrollment condition: at 3 seconds, LEAP obtains 0.652, exceeding the strongest outcome-free baseline, learned static ranking, by $+0.033$. The improvement remains at 5, 7, and 10 seconds, with gains of $+0.028$, $+0.023$, and $+0.017$, respectively. LEAP also exceeds the language-specific handcrafted baseline by $+0.040$, $+0.033$, $+0.027$, and $+0.020$ across the four budgets.

These results show that enrollment content matters most when only one or a few short utterances can be recorded; additional recording time reduces, but does not eliminate, the cost of suboptimal content selection. The comparison with learned static ranking isolates the value of set-conditioned selection. These methods use the same text-side evidence annotations and the same offline utility supervision, but the static ranker cannot represent the evidence already selected, candidate novelty or overlap, or the remaining recording budget. LEAP improves all three automatic components at every budget (Table 6), supporting the conclusion that its gain is not driven by a single metric. We examine this set-conditioned effect directly in Section 5.2. The advantage over learned static ranking is statistically reliable under all four pre-specified budget comparisons (Table 8). The standardized paired effect size decreases from 0.71 at 3 seconds to 0.41 at 10 seconds, consistent with selection decisions being most consequential at lower recording budgets. Realized-duration diagnostics show that LEAP operates under the same estimate-based feasibility rule as all baselines and does not obtain its gains through greater realized recording time (Table 7). Finally, the empirical oracle indicates that additional selection headroom remains, particularly at the 3-second budget. Because it accesses realized target-speaker outcomes during selection, however, it is not a deployable method. Detailed breakdown of performance gains across languages and components, actual budget utilization, and comparison with the strongest baseline are provided in Appendix C.1.

Table 2: **Active enrollment under matched estimated-duration budgets.** Values are normalized composite personalization utility (↑), averaged over Korean, Japanese, and Vietnamese. All methods use identical test speakers, downstream models, prompt inventories, and estimated-duration budgets. *Oracle* accesses downstream outcomes during selection and is non-deployable.

Method	3 s	5 s	7 s	10 s	Avg.
Random selection	0.556	0.604	0.642	0.679	0.620
Natural User selection	0.568	0.616	0.651	0.685	0.630
Phoneme coverage	0.593	0.638	0.670	0.698	0.650
Language handcrafted	0.612	0.654	0.682	0.707	0.664
Learned static ranking	0.619	0.659	0.686	0.710	0.669
LEAP (ours)	**0.652**	**0.687**	**0.709**	**0.727**	**0.694**
Oracle (*non-deployable*)	0.701	0.727	0.744	0.756	0.732

5.2 Dynamic Selection Outperforms Static Utterance Ranking

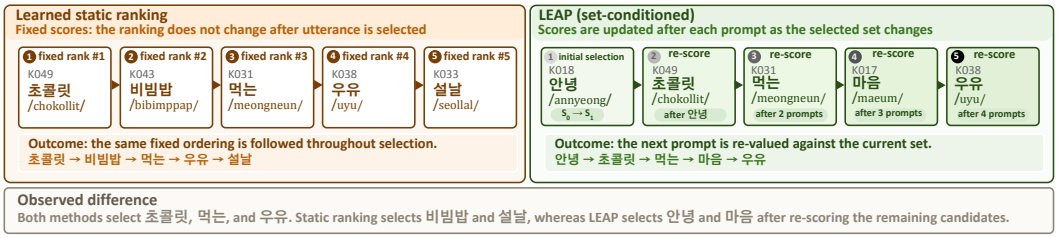


Figure 1: **Raw fixed-ranking prefix and set-conditioned selection trajectory from the Korean prompt inventory.** Both selectors observe the same candidate inventory. **Left:** learned static ranking follows an unchanged global ordering: *초콜릿* (*chokollit*; K049), *비빔밥* (*bibimppap*; K043), *먹는* (*meongneun*; K031), *우유* (*uyu*; K038), and *설날* (*seollal*; K033). **Right:** LEAP re-scores the remaining candidates after each selection, as indicated by the changing selected-set context, and chooses *안녕* (*annyeong*; K018), *초콜릿*, *먹는*, *마음* (*maeum*; K017), and *우유*, whose modeled duration totals 3.00 seconds under the 3-second budget. All quantitative results in Table 2 use feasible selection under matched recording budgets.

We next examine whether LEAP's improvement over learned static ranking arises from re-scoring candidates as enrollment proceeds. Learned static ranking uses the same text-side linguistic annotations and offline utility supervision as LEAP, but omits representations of the current enrollment set, candidate novelty and overlap, and the remaining budget. It therefore assigns each candidate a fixed score throughout selection. LEAP instead re-scores the remaining candidates relative to the evidence already selected and the available recording budget. Table 2 shows that learned static ranking improves over non-learned baselines, confirming that learned utterance-level preferences are useful. LEAP nevertheless outperforms it at every budget, with the largest gain at 3 seconds (+0.033). Thus, the improvement cannot be explained solely by identifying globally useful prompts: the value of an utterance depends on the evolving enrollment state. Figure 1 illustrates this distinction on a

representative Korean selection trajectory. Appendix C.8 confirms the pattern quantitatively: LEAP selects sets with broader canonical-descriptor coverage and lower redundancy, while the state–budget ablation shows that budget awareness alone does not account for the gain. Analogous set-conditioned trajectories for Japanese and Vietnamese are shown in Figure 3.

5.3 Enrollment Utility Transfers to Unseen Languages

We next test whether a policy trained on utility outcomes from one set of languages can select enrollment content for a language with no such outcomes. For each primary target language, LEAP and learned static ranking are trained only on the remaining source languages. The target contributes a prompt inventory and text-side linguistic descriptors, but no downstream utility outcomes, calibration statistics, policy updates, validation scores, or model-selection decisions.

Table 3: **Outcome-free transfer to unseen languages.** Normalized composite personalization utility (↑), averaged across estimated-duration budgets. LEAP and learned static ranking use no downstream outcomes from the target language. English is an additional transfer target trained using all three non-English source languages. † denotes a Holm-corrected paired speaker-level improvement over learned static ranking. **ko** (Korean), **ja** (Japanese), **vi** (Vietnamese), and **en** (English).

Target	Train	Phon. cov.	Handcraft	Static	LEAP
ko	**ja + vi**	0.602	0.618	0.624	**0.647**†
ja	**ko + vi**	0.617	0.631	0.636	**0.658**†
vi	**ko + ja**	0.629	0.646	0.651	**0.674**†
Additional transfer target					
en	**ko + ja + vi**	0.591	0.606	0.612	**0.635**†

Table 3 shows that LEAP transfers to every held-out primary language, improving over learned static ranking by $+0.022$ to $+0.023$ across Korean, Japanese, and Vietnamese. Because static ranking uses the same source-language outcomes and transferable descriptors, this gain shows the value of transferring a dynamic, state-aware selection policy rather than a fixed global prompt ranking. LEAP also exceeds the target-language handcrafted objective in every case. As an additional transfer target, LEAP attains 0.635 on English without using English utility outcomes, calibration statistics, validation scores, or model-selection signals during policy development. Thus, given only a target-language prompt inventory and text-side linguistic descriptors, the learned policy transfers beyond the languages from which utility labels were collected. Per-budget, component-level, and statistical results appear in Appendix C.2.

5.4 Human Listeners Prefer Actively Selected Enrollment

Automatic metrics provide a scalable measure of personalization quality, but do not by themselves establish whether listeners perceive the resulting voice as better matched to the target speaker. We therefore conduct a blinded paired listening study to test whether the automatic gains from active enrollment are also reflected in human preference. We recruit 54 valid native or highly proficient listeners, with 18 listeners per language. The study uses the pre-specified 3- and 7-second enrollment budgets, counterbalanced across trials. In each trial, listeners compare two converted utterances generated from the same source recording for the same target speaker; the systems differ only in the enrollment recordings selected for that speaker. For speaker-similarity trials, listeners additionally hear a held-out natural recording of the target speaker. Presentation order is randomized, and comparisons, target speakers, source utterances, and budgets are counterbalanced. Listeners choose which output better matches the target speaker, sounds more natural, or preserves the intended linguistic content more clearly.

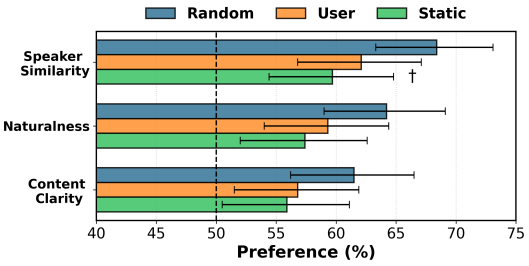


Figure 2: **Human preference for actively selected enrollment.** Bars show the percentage of blinded pairwise judgments favoring LEAP over the indicated baseline, pooled across Korean, Japanese, and Vietnamese. Error bars denote 95% confidence intervals; the dashed line indicates chance preference. † marks the pre-specified primary comparison, LEAP vs learned static ranking on speaker similarity.

Figure 2 summarizes pooled preferences across Korean, Japanese, and Vietnamese. Our pre-specified primary comparison is LEAP against learned static ranking on speaker similarity. Listeners choose LEAP in 59.7% of these judgments, consistent with the automatic advantage of dynamic, state-aware enrollment selection. We analyze this comparison using a mixed-effects logistic regression of the probability of choosing LEAP, with language and budget as fixed effects and listener, target speaker, and evaluation utterance as random effects. The remaining comparisons and dimensions are supporting analyses. They show the same directional preference for LEAP in naturalness and content clarity, with the largest margins against random selection. LEAP is also preferred to natural user-selected enrollment across all three dimensions, indicating that intuitively chosen prompts need not provide the most complementary evidence for downstream voice personalization. Per-language descriptives, confidence intervals, listener screening and quality-control analyses, and full statistical results are reported in Appendix C.3.

5.5 Ablations of Policy Design and Offline Utility Learning

We ablate the main policy-design and offline-utility-learning choices in LEAP using identical speaker-disjoint splits, prompt inventories, and downstream models. Table 4 shows that every component contributes to utility. Replacing set-conditioned scoring with set-independent scoring reduces average utility from 0.694 to 0.669, matching learned static ranking. Removing novelty and overlap features or duration-normalized acquisition also reduces performance across the evaluated budgets. Restricting the policy to segmental descriptors yields 0.668, showing that prosodic, structural, and transition-related evidence contributes beyond phoneme coverage alone. Finally, removing the listwise preference term and retaining only the regression loss reduces utility to 0.677, supporting the ranking objective for learning relative preferences among feasible next utterances. We additionally find that replacing the default 16-action supervision subsets with 32-action subsets or all feasible actions yields similar test utility, whereas collecting actions uniformly at random, without coverage-based and exploratory proposals, reduces performance. Full action-collection results are reported in Appendix C.4.

Table 4: **Core ablations.** Utility (↑) is averaged over Korean, Japanese, Vietnamese, and the four recording budgets.

Variant	3 s	Avg.
Full LEAP	**0.652**	**0.694**
set-independent scoring	0.619	0.669
without novelty and overlap	0.628	0.676
without duration-normalized	0.636	0.681
segmental descriptors only	0.622	0.668
regression loss only	0.631	0.677

6 Linguistic rationale.

LEAP's representation choice is consistent with the linguistic fact that an utterance is not characterized by its segmental inventory alone. Phoneme coverage is a useful proxy for articulatory variety, but it does not determine whether selected material represents the timing structure, prosodic contrasts, context-dependent realizations, or sequential dependencies through which segments are produced and perceived. These properties are complementary rather than interchangeable: for example, a segment may be present in the enrollment set while its coda, moraic, tonal, boundary, or neighboring-segment contexts remain unrepresented. This motivates representing prompts as configurations of segmental, structural, prosodic, and transition-related evidence, and scoring their marginal value relative to evidence already selected. The resulting policy is therefore not tied to a universal list of preferred words. Instead, it learns to favor complementary evidence configurations that can be instantiated by different lexical items in different languages. The cross-language and lexical-realization results are consistent with this interpretation: the transferable object is a preference for complementary linguistic evidence, not a memorized multilingual script.

7 Conclusion

We introduced LEAP, a language-transferable active-enrollment policy that sequentially selects complementary utterances under a recording-time budget using only text-side linguistic descriptors and estimated cost at deployment. Across matched budgets, LEAP outperforms random, user-selected, coverage-based, handcrafted, and learned static-ranking baselines, with the largest gains when recording time is scarce; listeners also prefer LEAP-selected enrollment. LEAP transfers across unseen speakers, budgets, lexical realizations, voice-personalization backends, and languages, offering a low-friction path to multilingual enrollment without collecting target-language utility outcomes.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript for grammar and readability. We have not used generative AI tools for any other tasks with required or recommended disclosures. We have reviewed all AI-assisted text, and we take full responsibility for the final content of this work.

REFERENCES

Zhiqi Ai, Cheng Han, Shiyi Mu, Zhiyong Chen, Yongjin Zhou, and Shugong Xu. Stabilizing Short Duration Speaker Verification through Neural Re-scoring with Hybrid Enrollment. In *Interspeech 2026*, pp. 6328–6332, 2026. doi: 10.21437/Interspeech.2026-228.

Baris Bozkurt, Ozlem Ozturk, and Thierry Dutoit. Text design for TTS speech corpus building using a modified greedy selection. In *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pp. 277–280, 2003. doi: 10.21437/Eurospeech.2003-118.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2709–2720. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/casanova22a.html.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pp. 4978–4982, 2024. doi: 10.21437/Interspeech.2024-2016.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16 (6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.313. URL https://aclanthology.org/2025.acl-long.313/.

Gaoxiang Cong, Yuankai Qi, Liang Li, Amin Beheshti, Zhedong Zhang, Anton van den Hengel, Ming-Hsuan Yang, Chenggang Yan, and Qingming Huang. Styledubber: Towards multi-scale style learning for movie dubbing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6767–6779, 2024.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report, 2026. URL https://arxiv.org/abs/2601.15621.

Shehzeen Samarah Hussain, Paarth Neekhara, Xuesong Yang, Edresson Casanova, Subhankar Ghosh, Roy Fejgin, Mikyas T. Desta, Rafael Valle, and Jason Li. Koel-TTS: Enhancing LLM based speech generation with preference alignment and classifier free guidance. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21219–21234, Suzhou,

China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1076. URL https://aclanthology.org/2025.emnlp-main.1076/.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/6832a7b24bc06775d02b7406880b93fc-Paper.pdf.

Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, Xiang Yin, Zejun MA, and Zhou Zhao. Mega-tts 2: Boosting prompting mechanisms for zero-shot speech synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 57919–57939, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/ff997469ac66cf893c4183efeb22212a-Paper-Conference.pdf.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22605–22623. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ju24b.html.

Ki-Joong Kwon, Junho So, and Sang-Hoon Lee. Coreaspeech: Korean speech corpus via jamo-based coreset selection for efficient and robust korean speech generation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-2114. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5b221bc040a287a255a0f311598aec61-Paper-Datasets_and_Benchmarks_Track.pdf.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 14005–14034. Curran Associates, Inc., 2023. doi: 10.52202/075280-0618. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2d8911db9ecedf866015091b28946e15-Paper-Conference.pdf.

Keon Lee, Dong Won Kim, Jaehyeon Kim, Seungjun Chung, and Jaewoong Cho. Ditto-tts: Diffusion transformers for scalable text-to-speech without domain-specific factors. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 52022–52055, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/80e77d9ed2f74dcaf1a42cb1a2593559-Paper-Conference.pdf.

Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095191.

Yinghao Aaron Li, Rithesh Kumar, and Zeyu Jin. DMOSpeech: Direct metric optimization via distilled diffusion model in zero-shot speech synthesis. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35186–35208. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25ay.html.

Songting Liu. Zero-shot voice conversion with diffusion transformers, 2024. URL https://arxiv.org/abs/2411.09943.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Gabriel Mittag and Sebastian Möller. Deep Learning Based Assessment of Synthetic Speech Naturalness. In *Interspeech 2020*, pp. 1748–1752, 2020. doi: 10.21437/Interspeech.2020-2382.

Takashi Nose, Yusuke Arao, Takao Kobayashi, Komei Sugiura, Yoshinori Shiga, and Akinori Ito. Entropy-based sentence selection for speech synthesis using phonetic and prosodic contexts. In *Interspeech 2015*, pp. 3491–3495, 2015. doi: 10.21437/Interspeech.2015-692.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, sheng zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 698–722, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/035a73893121b4534bb3314e831050b1-Paper-Conference.pdf.

Jan P. H. van Santen and Adam L. Buchsbaum. Methods for optimal text selection. In *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, pp. 553–556, 1997. doi: 10.21437/Eurospeech.1997-207.

Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Retrieval Augmented Generation in Prompt-based Text-to-Speech Synthesis with Context-Aware Contrastive Language-Audio Pretraining. In *Interspeech 2024*, pp. 1800–1804, 2024. doi: 10.21437/Interspeech.2024-980.

Hossein Zeinali, Lukás Burget, Hossein Sameti, and Honza Cernocký. Spoken pass-phrase verification in the i-vector space. In Anthony Larcher and Jean-François Bonastre (eds.), *Odyssey 2018: The Speaker and Language Recognition Workshop, 26-29 June 2018, Les Sables d'Olonne, France*, pp. 372–377. ISCA, 2018. doi: 10.21437/ODYSSEY.2018-52. URL https://doi.org/10.21437/Odyssey.2018-52.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, and Mingbo Ma. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 58436–58459, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9328208f88ec69420031647e6ff97727-Paper-Conference.pdf.

Zhisheng Zheng, Puyuan Peng, Anuj Diwan, Cong Phuoc Huynh, Xiaohang Sun, Zhu Liu, Vimal Bhat, and David Harwath. VoiceCraft-X: Unifying multilingual, voice-cloning speech synthesis and speech editing. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2737–2756, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.137. URL https://aclanthology.org/2025.emnlp-main.137/.

A METHOD DETAILS

A.1 CANDIDATE LEXICONS AND EVIDENCE ANNOTATIONS

Each candidate entry contains text, a pronunciation representation, estimated duration, and language-relevant linguistic annotations. We organize these annotations into the shared functional families $\mathcal{G} = \{\mathrm{SEG}, \mathrm{SYL}, \mathrm{PROS}, \mathrm{PROC}, \mathrm{CLST}, \mathrm{PTR}, \mathrm{BTR}, \mathrm{VIS}, \mathrm{VTR}, \mathrm{AVT}\}$. They represent, respectively: segmental units; syllable, mora, coda, and timing structure; prosodic contrasts; pronunciation processes; complex consonantal sequences; within-utterance transitions; cross-boundary transitions; visemic units; visemic transitions; and aligned phoneme–viseme transitions.

The lexicons preserve language-specific distinctions. Korean annotations include onset–nucleus–coda structure, segmental and boundary transitions, and morphophonological processes. Japanese annotations include moraic structure, special morae, long-vowel and geminate contrasts, phonological processes, and pitch-accent contrasts. Vietnamese annotations include onset–nucleus–coda structure, tone and tone-transition contrasts, and checked-syllable structure. English annotations include phonemic and surface-pronunciation forms, word-final codas, consonant clusters, stress, reduced forms, and pronunciation processes. Visemic and audiovisual-transition annotations are constructed for all languages.

For an annotated event $a \in A_g(x_i)$, we deterministically form a canonical descriptor $v_g(a)$ that captures its functional role. Segmental descriptors contain articulatory features such as place, manner, voicing, vowel height, backness, rounding, and length. Structural descriptors encode the unit type and structural position. Prosodic descriptors encode the contrast type, value, and position. Process descriptors encode functional classes such as assimilation, reduction, deletion, fortition, and lengthening, together with relevant conditioning context. Transition descriptors concatenate canonical descriptors for their source and destination events. Visemic and audiovisual descriptors use a shared visual-category inventory. These mappings expose comparable evidence without treating surface segment identities as shared across languages.

A.2 COMPLETE POLICY REPRESENTATION

For selection, coverage, novelty, and overlap are computed over unique canonical descriptors rather than raw event instances. Let

$$\mathcal{D}_g(x_i) = \{v_g(a) : a \in A_g(x_i)\}$$

denote the set of canonical descriptors instantiated by candidate x_i in family g. Thus, two language-specific events contribute the same evidence type whenever they receive the same functional descriptor, while repeated instances of a descriptor within an utterance count once for set-level coverage. A family-specific encoder maps each descriptor to a 64-dimensional embedding, $z_g(v) = E_g(v)$. For any descriptor set $\mathcal{D}$, let $\mathrm{Pool}(\mathcal{D})$ denote the concatenation of mean and max pooling over $\{z_g(v) : v \in \mathcal{D}\}$, with a learned family-specific null embedding for the empty set. The candidate and selected-set representations are

$$e_g(x_i) = \mathrm{Pool}(\mathcal{D}_g(x_i)), \qquad e_g(S_t) = \mathrm{Pool}\left(\bigcup_{x_j \in S_t} \mathcal{D}_g(x_j)\right).$$

We explicitly separate evidence introduced by a candidate from evidence already represented in the enrollment set:

$$\mathcal{D}_g^{\mathrm{new}}(x_i \mid S_t) = \mathcal{D}_g(x_i) \setminus \bigcup_{x_j \in S_t} \mathcal{D}_g(x_j), \qquad \mathcal{D}_g^{\mathrm{overlap}}(x_i \mid S_t) = \mathcal{D}_g(x_i) \cap \bigcup_{x_j \in S_t} \mathcal{D}_g(x_j).$$

The corresponding pooled representations are

$$e_g^{\mathrm{new}}(x_i \mid S_t) = \mathrm{Pool}(\mathcal{D}_g^{\mathrm{new}}(x_i \mid S_t)), \qquad e_g^{\mathrm{overlap}}(x_i \mid S_t) = \mathrm{Pool}(\mathcal{D}_g^{\mathrm{overlap}}(x_i \mid S_t)).$$

Let

$$\mathcal{V}_g^{\ell} = \bigcup_{x \in \mathcal{C}^{\ell}} \mathcal{D}_g(x), \qquad D_g^{\ell} = \max(1, |\mathcal{V}_g^{\ell}|).$$

We compute

$$c_g(S_t) = \frac{\left|\bigcup_{x_j \in S_t} \mathcal{D}_g(x_j)\right|}{D_g^\ell}, \qquad n_g(x_i \mid S_t) = \frac{|\mathcal{D}_g^{\text{new}}(x_i \mid S_t)|}{D_g^\ell},$$

$$r_g(x_i \mid S_t) = \frac{|\mathcal{D}_g^{\text{overlap}}(x_i \mid S_t)|}{\max(1, |\mathcal{D}_g(x_i)|)}.$$

For each family, the policy concatenates candidate, set, new, and overlap embeddings with these three scalars and the applicability mask $\rho_g^\ell \in \{0, 1\}$. Inapplicable families have all associated embeddings and scalar statistics set to zero. The full input additionally includes normalized syllable or mora count, pronunciation-process count, within-utterance transition count, boundary-transition count, relative duration, remaining budget, and set size. Each count is divided by the maximum value of the corresponding feature over $\mathcal{C}^\ell$, with denominator clipped below by one. A shared multilayer scorer f_θ receives this concatenated vector. No learned language, speaker, or downstream-model identity is provided.

A.3 Offline Utility Learning

For a speaker s, source model m, state S_t, and candidate addition x_i, we obtain the empirical marginal-utility target

$$y_{i,t}^{s,m} = Q(S_t \cup \{x_i\}; s, m) - Q(S_t; s, m).$$

The utility Q is a fixed weighted sum of automatic speaker-similarity, content-preservation, and predicted-naturalness measures. All component metrics are oriented so that higher is better and standardized using training speakers only. To make source language–model conditions comparable, marginal-utility targets are standardized within their corresponding source condition:

$$\widetilde{y}_{i,t}^{s,m} = \frac{y_{i,t}^{s,m} - \mu_{\ell,m}}{\sigma_{\ell,m} + \epsilon}.$$

Here, $\mu_{\ell,m}$ and $\sigma_{\ell,m}$ are computed once from all marginal-utility targets collected from training speakers in source language–model condition (ℓ, m). They are frozen before validation and test evaluation and are never computed from held-out speakers, languages, or voice-personalization models.

We collect state–candidate tuples from random, coverage-based, and exploratory trajectories at multiple budgets. At each visited state, we evaluate a bounded action subset $\mathcal{A}_t$ containing randomly sampled candidates and candidates proposed by the coverage-based and exploratory policies. Outputs and evaluations for recurrent enrollment sets are cached and reused.

For each evaluated action set, the target and predicted preference distributions are

$$\pi_{i,t}^* = \frac{\exp(\widetilde{y}_{i,t}^{s,m}/\tau)}{\sum_{x_j \in \mathcal{A}_t} \exp(\widetilde{y}_{j,t}^{s,m}/\tau)}, \qquad \pi_{i,t}^\theta = \frac{\exp(\hat{u}_{i,t}/\tau)}{\sum_{x_j \in \mathcal{A}_t} \exp(\hat{u}_{j,t}/\tau)}.$$

Let $\mathcal{T}_{\text{train}}$ denote all sampled speaker–model–state tuples (s, m, t) in the training collection. We optimize

$$\mathcal{L} = -\sum_{(s,m,t) \in \mathcal{T}_{\text{train}}} \sum_{x_i \in \mathcal{A}_t} \pi_{i,t}^* \log \pi_{i,t}^\theta + \lambda \sum_{(s,m,t) \in \mathcal{T}_{\text{train}}} \sum_{x_i \in \mathcal{A}_t} \text{Huber}(\hat{u}_{i,t}, \widetilde{y}_{i,t}^{s,m}).$$

Training trajectories, normalization statistics, early stopping, and hyperparameter selection use training speakers and source conditions only.

A.4 Budget-Constrained Selection Algorithm

The policy uses estimated durations only for feasibility and duration-normalized acquisition. Realized durations are measured separately under the matched-duration evaluation protocol.

Algorithm 1 LEAP budget-constrained enrollment selection

Require: candidate inventory $\mathcal{C}^\ell$, annotations A, estimated durations $\{\hat{d}_i\}$, budget B, scorer f_θ, exponent γ

$S \leftarrow \varnothing$; $b \leftarrow B$
while $\exists x_i \in \mathcal{C}^\ell \setminus S$ such that $\hat{d}_i \le b$ **do**
  $\mathcal{F} \leftarrow \{x_i \in \mathcal{C}^\ell \setminus S : \hat{d}_i \le b\}$
  for all $x_i \in \mathcal{F}$ **do**
    construct h_i from $A(x_i)$, $A(S)$, and b
    $a_i \leftarrow f_\theta(h_i)/\hat{d}_i^{\gamma}$
  end for
  $x^* \leftarrow \arg\max_{x_i \in \mathcal{F}} a_i$
  $S \leftarrow S \cup \{x^*\}$; $b \leftarrow b - \hat{d}_{x^*}$
end while
return S

B Experimental Setup Details

B.1 Lexicon Construction and Recording Protocol

Each Korean, Japanese, and Vietnamese candidate lexicon contains 96 prompts spanning short lexical items and sentence-level prompts. Every entry includes text, pronunciation information, an estimated duration, and annotations for segmental units; syllable, mora, coda, and timing structure; prosodic contrasts; pronunciation processes; consonant clusters where applicable; within-utterance and cross-boundary transitions; visemes; visual transitions; and aligned audiovisual events. We remove duplicate prompts and entries with invalid or incomplete pronunciation annotations. The language-specific mapping to the shared evidence families is described in Appendix A.1.

B.1.1 Recording Protocol

All speakers record the full lexicon in their native language under a common protocol at 24 kHz. We apply the same silence trimming, loudness normalization, and quality-control procedure to every recording. Clips with incorrect readings, truncation, substantial disfluency, or excessive background noise are rerecorded or excluded. The train/validation/test split is fixed before policy training and stratified by gender where possible. For each primary language, we recruit 30 native speakers and use mutually exclusive 18/6/6 train/validation/test splits, stratified by gender where possible.

For every primary-language test speaker, we use the same 20 disjoint source utterances for every enrollment method, model, and recording budget. Each target speaker additionally provides five disjoint natural recordings used only to construct the speaker-similarity centroid; these recordings are not used as enrollment candidates or source utterances.

English outcome-free transfer cohort. English is used only as an outcome-free transfer target. We construct an English candidate lexicon using the same text-side annotation pipeline and collect a speaker-disjoint English evaluation cohort before running any selection experiment. The cohort is split into an enrollment/evaluation pool and a disjoint source-speech pool; no English recording, personalized output, normalization statistic, or validation score is used to train, tune, or select the policy. Each English target speaker records the full 96-prompt lexicon and five additional natural utterances used only to form the speaker-similarity centroid. For each target speaker, the same 20 disjoint source utterances and the same 3-, 5-, 7-, and 10-second estimated-budget protocol are used for every selection method.

Recording provenance and quality control. Before analysis, we freeze a recording manifest containing recruitment source, native-language eligibility, self-reported demographic fields, compensation, consent, and the recording setup (microphone or device, room condition, sampling rate, and capture software). We report aggregate demographics only. The manifest also records every initial clip, rerecording request, final retained clip, and exclusion reason. A clip is rerecorded when it contains an incorrect prompt reading, truncation, substantial disfluency, clipping, or excessive background noise;

a clip that remains unusable after the allowed rerecording attempts is excluded before splitting. Thus, exclusions are based on recording quality rather than any downstream personalization metric.

B.2 Voice-Model and Budget Details

Seed-VC (Liu, 2024) is a diffusion-based voice-conversion model that extracts linguistic content with Whisper and synthesizes speech with a conditional diffusion model and BigVGAN vocoder. We use the official Seed-VC checkpoint `DiT_seed_v2_uvit_whisper_small_wavenet_bigvgan_pruned.pth` with 25 diffusion steps, classifier-free guidance 0.7, no explicit F_0 conditioning, and no automatic pitch adjustment.

OpenVoice V2 (Qin et al., 2023) transfers target-speaker tone color separately from linguistic content. We use the official converter checkpoint and configuration, disable additional style manipulation, and use the system only for zero-shot tone-color conversion.

FreeVC (Li et al., 2023) is a text-free one-shot voice-conversion model based on WavLM (Chen et al., 2022) content representations; we use the standard FreeVC model, rather than the speaker-embedding-only FreeVC-s variant, with the official `freevc.pth`, `WavLM-Large.pt`, and configuration.

All checkpoints remain frozen. For models accepting a single reference file, we concatenate selected enrollment clips in fixed candidate-identifier order, separated by 100 ms silence. Source speakers are disjoint from benchmark target speakers and never serve as target references. We use the same source utterances, random seeds, and preprocessing for every enrollment method.

Reproducibility manifest and model failures. Before the test runs, we freeze a machine-readable manifest for every model containing the repository URL, commit or release tag, checkpoint filename and SHA-256 hash, license, dependency environment, hardware, accelerator count, and runtime version. The manifest records the input sampling rate, each resampling operation, random seed, and inference configuration. We define a model failure before examining outcomes as an unrecoverable runtime error, non-finite waveform, empty waveform, duration below 0.25 seconds, or clipping in more than 1% of samples. Failures are rerun once with the identical frozen configuration; persistent failures are reported by model, language, budget, and method, and are excluded from all methods in the affected matched cell.

B.2.1 Budget Handling

Candidate feasibility is determined by $\hat{d}_i$ only. If no candidate fits the initial budget, the enrollment set is empty and the case is reported as infeasible. Otherwise, selection continues until no feasible candidate remains, even if all remaining predicted utilities are negative. We retain realized duration above the nominal budget rather than truncating recordings, because this deviation is not observable at selection time. For cross-budget transfer, a policy is trained on states from three of $\{3, 5, 7, 10\}$ seconds and evaluated on the excluded budget without fine-tuning.

B.3 Baselines and Policy Supervision

Uniform random selection samples feasible candidates without replacement until no candidate fits; we use 10 independently sampled enrollment sets for each speaker, language, model, and budget. Natural user-selected enrollment presents the candidate pool without annotations, coverage statistics, utilities, or recommendations and applies the same estimated-duration constraint.

Phoneme coverage greedily selects

$$x_t^* = \arg\max_{x_i \in \mathcal{F}_t} \frac{\left|\mathcal{D}_{\mathrm{SEG}}(x_i) \setminus \bigcup_{x \in S_t} \mathcal{D}_{\mathrm{SEG}}(x)\right|}{\hat{d}_i}.$$

The handcrafted baseline combines prespecified, applicable language-specific coverage components with uniform weights. At each step, it scores a feasible candidate by the uniformly weighted sum of newly covered canonical descriptors across the components applicable to the language, divided by its estimated duration. The learned static ranker uses the same text-side annotations as LEAP but omits

current-set, novelty, overlap, and remaining-budget features. It is trained with the same offline utility targets, action sets, and objective as LEAP, but its score depends only on candidate-side features and relative duration.

The oracle instead observes test-speaker outcomes and greedily selects

$$x_t^{\text{oracle}} = \arg\max_{x_i \in \mathcal{F}_t} \frac{Q(S_t \cup \{x_i\}; s, m) - Q(S_t; s, m)}{\hat{d}_i}.$$

It is non-deployable and used only as a headroom reference. All deterministic methods break exact ties by candidate identifier.

B.3.1 Policy Supervision

We define

$$Q(S; s, m) = \tfrac{1}{3}\widetilde{M}_{\text{spk}}(S; s, m) + \tfrac{1}{3}\widetilde{M}_{\text{int}}(S; s, m) + \tfrac{1}{3}\widetilde{M}_{\text{nat}}(S; s, m),$$

where all components are oriented so that higher is better and standardized using training-speaker outcomes in the corresponding language–model condition. The empirical marginal target is $y_{i,t}^{s,m} = Q(S_t \cup \{x_i\}; s, m) - Q(S_t; s, m)$; marginal labels are subsequently standardized within each source language–model condition.

We collect supervision in four iterative rounds. The first uses uniform-random, phoneme-coverage, and handcrafted-coverage trajectories. Later rounds add a frozen exploratory policy trained on all previous-round supervision. At each state, we evaluate at most 16 actions: eight uniformly sampled candidates, four coverage-policy proposals, and four exploratory-policy proposals; duplicate proposals are replaced by uniform samples. If fewer than 16 actions remain feasible, we evaluate all of them. Ranking losses are computed only among actions evaluated in the same state. We cache recurrent enrollment sets, model inputs, target representations, and converted outputs. We assess sensitivity using action-set sizes 16 and 32 and all feasible actions on sampled validation states.

The scorer uses 64-dimensional event encoders, a three-layer MLP with hidden dimension 256, GELU activations, and dropout 0.1. We optimize with AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}), batch size 128, gradient clipping at 1.0, and at most 100 epochs. Early stopping uses validation NDCG@10 with patience 10. We set $\tau = 0.5$, $\lambda = 0.5$, and $\gamma = 1.0$.

B.4 Transfer, Metrics, and Human Evaluation

In every leave-one-language-out fold, target-language recordings, outcomes, normalization statistics, validation scores, and outcome-derived rankings are excluded from development. For English transfer, English text-side annotations, candidate inventories, and estimated durations are available at inference, but no English outcome is used for training or model selection. In every leave-one-model-out fold, outputs, utility labels, normalization statistics, and validation signals from the held-out model are excluded.

Pre-specified strongest-baseline rule. We choose the human-comparison baseline before accessing any test-speaker or human-evaluation result. Let $\mathcal{B}$ contain all deployable non-oracle baselines: random selection, natural user selection, phoneme coverage, language-specific handcrafted utility, and the learned static ranker. For a standard within-language setting (ℓ, m, b), we select

$$b^*(\ell, m, b) = \arg\max_{b' \in \mathcal{B}} \frac{1}{|\mathcal{V}_\ell|} \sum_{s \in \mathcal{V}_\ell} Q_{\text{val}}(S_{b'}; s, m, b),$$

using only validation speakers $\mathcal{V}_\ell$ and the utility definition fixed in Section B.3.1. For outcome-free transfer to a held-out language, the target language is excluded from this choice: b^* is selected by the unweighted mean of the same validation utility over the source-language validation speakers, source models, and evaluated budgets. The selected baseline identity is frozen in a protocol file before test-waveform generation and is then used for all paired human comparisons in that setting. Exact ties are broken by the fixed baseline order stated above. This rule prevents post-hoc selection of a weak comparator while preserving the outcome-free status of target-language transfer.

B.4.1 AUTOMATIC METRICS

Speaker similarity is the cosine similarity between the converted-utterance embedding and a target-speaker centroid formed by averaging five disjoint natural recordings. We use frozen ECAPA–TDNN (Desplanques et al., 2020) embeddings after resampling to 16 kHz. Content preservation is character error rate from frozen Whisper large-v3 (Radford et al., 2023) in transcription mode with the correct language token, temperature zero, and no reference-text prompt; word error rate is additionally reported where reliable language-specific tokenization exists. Naturalness uses frozen NISQA–TTS (Mittag & Möller, 2020), which receives only the converted waveform. Scores are averaged over utterances within speaker, then macro-averaged across speakers. Confidence intervals use 10,000 hierarchical paired-bootstrap resamples.

B.4.2 HUMAN EVALUATION

Human evaluation uses all six test speakers in each primary language at 3 and 7 seconds. For each speaker and budget, five held-out source utterances are sampled, and LEAP is paired with the strongest deployable non-oracle baseline for that condition. Each pair receives five independent judgments from native or highly proficient listeners, with at least 18 retained listeners per language. System order is randomized and listeners are blind to method. They judge speaker identity, naturalness, and content clarity, with a target-speaker natural recording supplied for speaker-similarity judgments. They may prefer the first system, second system, or neither.

Participants complete four attention checks. We exclude participants failing at least two checks or completing the study in less than one third of the median completion time. For preference analysis, ties are reported separately and excluded from a mixed-effects logistic regression with method as a fixed effect and listener, target speaker, and source utterance as random intercepts. We apply Holm–Bonferroni correction within metric families.

Ethics, participant sampling, and counterbalancing. The study was approved by the Institutional Review Board (Exemption Number: PIRB-2025-E020, Date: June 9, 2025). We recruit adult listener and speaker through the local country universities. Participants must self-report native or high proficiency in the evaluated language and pass a language-appropriate instruction-comprehension check before starting. We report the number recruited, retained, and excluded by language, with every exclusion assigned to the pre-specified attention or completion-time rule above.

For each language, speaker, budget, source utterance, and judgment dimension, we present the LEAP–b^* pair in both system orders across listeners. Assignment uses balanced incomplete blocks so that each listener sees no source utterance more than once for a given dimension and each pair receives the same number of judgments in each order. We report the resulting total number of trials, the number of unique pairs, judgments per pair, and the realized order balance. The 3- and 7-second budgets, three judgment dimensions, eligibility/exclusion rules, and baseline-selection rule are frozen before recruitment

C RESULTS DETAILS

C.1 DETAILED PERFORMANCE ANALYSIS OF ACTIVE ENROLLMENT

Table 5 reports the normalized composite utility for each primary language. Tables 6, 7, and 8 provide the component metrics, budget diagnostics, and confirmatory statistical comparisons, respectively.

C.2 DETAILED RESULTS FOR OUTCOME-FREE LANGUAGE TRANSFER

Outcome-free normalization. For a leave-one-language-out target $\ell^\star$, metric k, and voice-personalization model m, we compute the evaluation calibration parameters

$$\mu_{k,m}^{-\ell^\star},\ \sigma_{k,m}^{-\ell^\star}$$

Table 5: **Per-language normalized composite personalization utility (↑).** Values are averaged over held-out speakers and downstream voice-personalization models within each language. The final six rows reproduce the unweighted macro-average reported in Table 2.

Language	Method	3 s	5 s	7 s	10 s	Avg.
Korean	Random selection	0.561	0.609	0.646	0.683	0.625
	Natural User selection	0.573	0.621	0.655	0.689	0.635
	Phoneme coverage	0.599	0.644	0.676	0.704	0.656
	Language handcrafted	0.618	0.660	0.688	0.713	0.670
	Learned static ranking	0.625	0.665	0.692	0.716	0.675
	LEAP (ours)	**0.659**	**0.694**	**0.716**	**0.734**	**0.701**
Japanese	Random selection	0.553	0.601	0.639	0.676	0.617
	Natural User selection	0.565	0.613	0.648	0.683	0.627
	Phoneme coverage	0.590	0.635	0.667	0.695	0.647
	Language handcrafted	0.607	0.649	0.678	0.703	0.659
	Learned static ranking	0.614	0.654	0.681	0.706	0.664
	LEAP (ours)	**0.644**	**0.679**	**0.702**	**0.720**	**0.686**
Vietnamese	Random selection	0.554	0.602	0.641	0.678	0.619
	Natural User selection	0.566	0.614	0.650	0.683	0.628
	Phoneme coverage	0.590	0.635	0.667	0.695	0.647
	Language handcrafted	0.611	0.653	0.680	0.705	0.662
	Learned static ranking	0.618	0.658	0.685	0.708	0.667
	LEAP (ours)	**0.653**	**0.688**	**0.709**	**0.727**	**0.694**
Macro-average	Random selection	0.556	0.604	0.642	0.679	0.620
	Natural User selection	0.568	0.616	0.651	0.685	0.630
	Phoneme coverage	0.593	0.638	0.670	0.698	0.650
	Language handcrafted	0.612	0.654	0.682	0.707	0.664
	Learned static ranking	0.619	0.659	0.686	0.710	0.669
	LEAP (ours)	**0.652**	**0.687**	**0.709**	**0.727**	**0.694**

Table 6: **Component-wise automatic metrics.** Scores are normalized within language–model conditions before macro-averaging. The composite is the unweighted mean of the three normalized component scores and matches Table 2. The learned static ranker is the strongest outcome-free baseline.

Budget	Method	Speaker sim.	Content pres.	Naturalness	Composite
3 s	Learned static ranking	0.643	0.614	0.600	0.619
	LEAP (ours)	**0.674**	**0.646**	**0.636**	**0.652**
5 s	Learned static ranking	0.681	0.654	0.642	0.659
	LEAP (ours)	**0.708**	**0.680**	**0.673**	**0.687**
7 s	Learned static ranking	0.705	0.680	0.673	0.686
	LEAP (ours)	**0.727**	**0.703**	**0.697**	**0.709**
10 s	Learned static ranking	0.727	0.704	0.699	0.710
	LEAP (ours)	**0.742**	**0.721**	**0.718**	**0.727**

from the pooled outcomes of training speakers in the source languages only. We then apply the frozen model-specific affine transformation

$$\widetilde{M}_k^{-\ell^\star}(S; s, m) = \frac{M_k(S; s, m) - \mu_{k,m}^{-\ell^\star}}{\sigma_{k,m}^{-\ell^\star} + \epsilon}$$

unchanged when evaluating target-language speakers. This source-only evaluation calibration is distinct from the source-condition target standardization used during offline utility learning (Appendix A.3). Thus, target-language outcomes do not contribute to composite normalization, metric calibration, policy training, early stopping, checkpoint selection, or model selection. This protocol applies equally to the English transfer setting, for which all three primary languages are sources.

Table 7: **Realized-duration diagnostics.** Selectors observe only estimated duration. Realized duration, overrun, and signed deviation are computed after recording and are not used for selection. Values are averaged over languages, held-out speakers, and voice-personalization models.

	3 s budget			5 s budget			7 s budget			10 s budget		
Method	**Real.**	**Over.**	**Δ**	**Real.**	**Over.**	**Δ**	**Real.**	**Over.**	**Δ**	**Real.**	**Over.**	**Δ**
Random selection	2.91	14.2%	-0.09	4.86	15.6%	-0.14	6.88	16.1%	-0.12	9.83	14.4%	-0.17
Natural user	2.89	13.9%	-0.11	4.84	15.0%	-0.16	6.85	15.8%	-0.15	9.80	14.1%	-0.20
Phoneme coverage	2.94	15.1%	-0.06	4.90	16.2%	-0.10	6.92	16.7%	-0.08	9.87	15.0%	-0.13
Language handcrafted	2.95	15.4%	-0.05	4.91	16.3%	-0.09	6.93	16.9%	-0.07	9.88	15.3%	-0.12
Learned static ranking	2.96	15.7%	-0.04	4.92	16.6%	-0.08	6.94	17.1%	-0.06	9.89	15.5%	-0.11
LEAP	2.95	15.5%	-0.05	4.92	16.4%	-0.08	6.94	17.0%	-0.06	9.89	15.4%	-0.11

Note: Real. denotes mean realized duration in seconds; Over. denotes the proportion of enrollment sets whose realized duration exceeds the nominal budget; $\Delta = d^{\mathrm{real}}(S) - B$ is the signed deviation in seconds.

Table 8: **Confirmatory comparison with the strongest outcome-free baseline.** Paired speaker-level bootstrap tests compare LEAP with learned static ranking. Confidence intervals and standardized paired effect sizes are computed over held-out speakers; p_{Holm} corrects the four pre-specified budget comparisons.

Budget	**Δ utility**	**95% CI**	**Effect size**	$\boldsymbol{p_{\mathrm{Holm}}}$
3 s	+0.033	[0.024, 0.042]	0.71	< 0.001
5 s	+0.028	[0.019, 0.037]	0.62	< 0.001
7 s	+0.023	[0.015, 0.032]	0.54	< 0.001
10 s	+0.017	[0.009, 0.026]	0.41	0.002

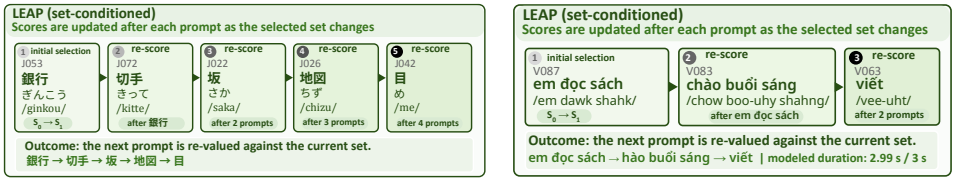
C.3 Human Evaluation Details

Participants, consent, and ethics. We recruited 61 listeners and retained 54 valid participants: 18 each for Korean, Japanese, and Vietnamese. Participants were native speakers of the evaluated language or self-reported highly proficient listeners who had used the language for at least five years and passed a language-comprehension screening item. All participants provided informed consent before taking part and were compensated for their time. The study was approved by the Institutional Review Board (Exemption Number: PIRB-2025-E020, Date: June 9, 2025).

Trial design and presentation. Each retained listener completed 72 blinded pairwise trials: 3 baseline comparisons (random selection, natural user-selected enrollment, and learned static ranking) $\times$ 3 judgment dimensions (speaker similarity, naturalness, and content clarity) $\times$ 2 enrollment budgets (3 and 7 seconds) $\times$ 4 sampled trials per cell. This yields 3,888 retained judgments overall and 432 judgments for each comparison–dimension cell.

For each language, trials draw from its six held-out target speakers and the same pool of 20 disjoint source utterances used for automatic evaluation. Within each listener, the four target-speaker–source-utterance pairs in a comparison–dimension–budget cell are sampled without replacement when feasible. Across listeners, target speakers, source utterances, budgets, and method-presentation order are balanced as closely as possible. The two outputs in each trial are generated from the same target speaker, source utterance, voice-personalization model, and budget; they differ only in the enrollment set. For speaker-similarity trials, a held-out natural recording of the target speaker is presented before the two synthesized outputs. Audio files are presented through an embedded player at a fixed normalized loudness; replay is permitted, but listeners must hear both outputs before responding. Left–right order is independently randomized for every trial.

Confirmatory analysis. The preregistered primary analysis fits a mixed-effects logistic regression to the speaker-similarity judgments comparing LEAP with learned static ranking. The dependent variable is whether the listener selects LEAP. The model includes language and budget as fixed effects and random intercepts for listener, target speaker, and source utterance. Supporting comparisons and secondary judgment dimensions are not used to select the primary analysis.



(a) Japanese **(b) Vietnamese**

Figure 3: **Set-conditioned selection trajectories in additional languages. (a) Japanese:** LEAP sequentially selects 銀行 (*ginkou*; J053), 切手 (*kitte*; J072), 坂 (*saka*; J022), 地図 (*chizu*; J026), and 目 (*me*; J042) under the 3-second recording budget. **(b) Vietnamese:** LEAP sequentially selects *em đọc sách* (*em dawk shahk*; V087), *chào buổi sáng* (*chow boo-uhy shahng*; V083), and *viết* (*vee-uht*; V063) under the same budget. In both languages, LEAP re-scores the remaining candidates after each selection conditioned on the current enrollment set.

Table 9: **Per-budget outcome-free transfer to unseen languages.** Normalized composite personalization utility (↑). Each target is evaluated with no target-language utility outcomes during policy development; the average reproduces Table 3.

Target	Method	3 s	5 s	7 s	10 s	Avg.
ko	Phoneme coverage	0.570	0.599	0.616	0.623	0.602
	Language handcrafted	0.586	0.615	0.632	0.639	0.618
	Learned static ranking	0.590	0.620	0.637	0.649	0.624
	LEAP	**0.615**	**0.645**	**0.661**	**0.667**	**0.647**
ja	Phoneme coverage	0.587	0.613	0.631	0.637	0.617
	Language handcrafted	0.600	0.627	0.645	0.652	0.631
	Learned static ranking	0.603	0.631	0.650	0.660	0.636
	LEAP	**0.628**	**0.654**	**0.671**	**0.679**	**0.658**
vi	Phoneme coverage	0.598	0.625	0.644	0.649	0.629
	Language handcrafted	0.616	0.643	0.660	0.665	0.646
	Learned static ranking	0.620	0.646	0.665	0.673	0.651
	LEAP	**0.643**	**0.670**	**0.688**	**0.695**	**0.674**
en	Phoneme coverage	0.562	0.586	0.605	0.611	0.591
	Language handcrafted	0.576	0.602	0.620	0.626	0.606
	Learned static ranking	0.580	0.606	0.626	0.636	0.612
	LEAP	**0.603**	**0.630**	**0.650**	**0.657**	**0.635**

C.4 Robustness to Offline Action Collection

The action subsets used to obtain offline marginal-utility labels affect only *training supervision collection*. They do not restrict deployment: at test time, LEAP scores every candidate in the feasible set $\mathcal{F}_t$ and selects the highest-scoring feasible candidate.

Our default collection protocol evaluates at most 16 actions per enrollment state: eight candidates sampled uniformly from $\mathcal{F}_t$, four proposals from the coverage-based policies, and four proposals from the frozen exploratory policy trained on preceding collection rounds. Duplicate proposals are removed and replaced by newly sampled uniform feasible candidates. If fewer than 16 actions are feasible, we evaluate all feasible actions.

For the 32-action variant, we evaluate 16 uniformly sampled candidates, eight coverage-policy proposals, and eight exploratory-policy proposals, again removing duplicates and replenishing them with uniform feasible candidates. For the all-feasible variant, we evaluate every candidate in $\mathcal{F}_t$ at each sampled state and use no proposal subsampling. The uniform-only variant retains the 16-action budget but removes both coverage-based and exploratory proposals: all 16 actions are sampled uniformly without replacement from $\mathcal{F}_t$. Thus, this variant removes targeted supervision of high-coverage and policy-relevant actions while leaving the model architecture, loss, and total action-label budget unchanged.

Table 10: **Component-wise outcome-free transfer results.** Metrics are source-normalized as described above and averaged over estimated recording budgets.

Target	Method	Speaker sim.	Content pres.	Naturalness	Composite
ko	Static	0.638	0.620	0.614	0.624
	LEAP	**0.660**	**0.647**	**0.634**	**0.647**
ja	Static	0.650	0.631	0.627	0.636
	LEAP	**0.672**	**0.657**	**0.645**	**0.658**
vi	Static	0.663	0.649	0.641	0.651
	LEAP	**0.687**	**0.675**	**0.660**	**0.674**
en	Static	0.625	0.609	0.602	0.612
	LEAP	**0.649**	**0.636**	**0.620**	**0.635**

Table 11: **Statistical comparison for outcome-free language transfer.** Paired speaker-level bootstrap comparisons test LEAP against learned static ranking. Confidence intervals are for the composite-utility difference; p_{Holm} corrects the four pre-specified target-language tests.

Target	Δ utility	95% CI	Effect size	p_{Holm}
ko	+0.023	[0.014, 0.032]	0.58	0.001
ja	+0.022	[0.013, 0.031]	0.55	0.002
vi	+0.023	[0.015, 0.032]	0.61	< 0.001
en	+0.023	[0.011, 0.035]	0.49	0.004

Increasing the number of evaluated actions from 16 to 32, or evaluating all feasible actions, yields results within seed variation of the default protocol. In contrast, uniform-only collection reduces utility, especially at the 3-second budget. Coverage-based and exploratory proposals therefore improve the informativeness of the offline labels without imposing a restricted candidate set at inference time.

C.5 Generalization to Unseen Speakers, Budgets, and Lexical Realizations

A practical active-enrollment policy must transfer beyond the speakers, recording budgets, and surface prompt realizations observed during offline utility learning. LEAP is not designed to memorize a fixed multilingual lexicon. Instead, it operates on linguistic evidence descriptors, allowing the same category specification to be instantiated with alternative eligible words or phrases while using the same trained policy. We evaluate generalization along three axes. First, all evaluations use test speakers disjoint from those used to obtain utility-learning outcomes. Second, we train and select the policy without outcomes from a target recording budget, then evaluate directly at that held-out budget. Third, we construct alternative prompt inventories by replacing surface lexical realizations while preserving the relevant language-specific linguistic categories. These test prompts do not occur in the inventory used for offline utility learning.

Table 17 first confirms that LEAP transfers to speakers whose downstream personalization outcomes were unavailable during utility learning. Consistent with Table 2, it improves over learned static ranking on speaker-disjoint test speakers by $+0.033$ at 3 seconds and $+0.017$ at 10 seconds. Thus, deployment does not require speaker-specific outcome measurements: LEAP selects enrollment content from text-side linguistic evidence alone. LEAP also transfers to recording budgets that were not observed during policy training or model selection. When 5-second and 10-second budgets are held out, LEAP outperforms learned static ranking by $+0.026$ and $+0.016$, respectively. The smaller gap at the longer budget is consistent with the main results: with more available recording time, a fixed ranking can include a broader range of useful prompts. Nevertheless, the positive improvement at held-out budgets shows that LEAP learns a duration-aware selection principle, rather than a separate policy tied to each predefined recording constraint. Finally, LEAP generalizes to unseen lexical realizations of the same prompt specifications. Replacing surface words and phrases with alternative eligible realizations leaves LEAP ahead of static ranking by $+0.029$ at 3 seconds and $+0.019$ at 10 seconds. Because these test prompts are absent from the utility-learning inventory while preserving the underlying linguistic categories, this result rules out an explanation based on memorizing a

Table 12: **Participant recruitment and quality control.**

Item	Count
Initially recruited listeners	61
Excluded: incomplete study	1
Excluded: failed headphone check	2
Excluded: failed attention checks	4
Final valid listeners	54
Valid listeners per language	18

Table 13: **Pooled human preference for LEAP over each baseline.** Values are percentages of blinded pairwise judgments favoring LEAP, with nonparametric 95% confidence intervals over judgments. Each row contains 432 retained judgments pooled across Korean, Japanese, and Vietnamese.

Baseline	Dimension	LEAP preference	95% CI	n
Random selection	Speaker similarity	68.4%	[64.0, 72.5]	432
	Naturalness	64.2%	[59.7, 68.5]	432
	Content clarity	61.5%	[56.9, 65.9]	432
Natural user-selected enrollment	Speaker similarity	62.1%	[57.6, 66.4]	432
	Naturalness	59.3%	[54.7, 63.7]	432
	Content clarity	56.8%	[52.2, 61.2]	432
Learned static ranking	Speaker similarity	59.7%	[55.1, 64.1]	432
	Naturalness	57.4%	[52.8, 61.8]	432
	Content clarity	55.9%	[51.3, 60.4]	432

fixed lexicon. Instead, LEAP transfers the utility of complementary linguistic evidence across valid re-instantiations of the enrollment inventory. Per-language results, additional resampling seeds, and component-wise metrics appear in Table 18.

Robustness to action-collection resampling. We repeat policy training with independently resampled offline trajectory–action collections while keeping speaker partitions, candidate inventories, and evaluation items fixed. Table 19 reports mean and standard deviation across resampling seeds. The relative advantage of LEAP persists under each generalization axis.

C.6 LEAP Transfers Across Voice-Personalization Models

We next test whether LEAP learns a selection policy tied to one particular voice-personalization backend or a transferable notion of complementary enrollment utility. This distinction matters because the three systems in our evaluation—Seed-VC, OpenVoice V2, and FreeVC—use substantially different mechanisms to represent linguistic content and target-speaker identity.

For each target model, we train LEAP using offline utility outcomes from the other two models only. The held-out target model contributes no utility-learning outcomes, calibration statistics, checkpoint-selection signals, or policy updates. At test time, all methods select from the same target-language prompt inventories under matched recording budgets, and the selected enrollment recordings are passed unchanged to the held-out model. Thus, this leave-one-model-out protocol evaluates whether active-enrollment utility transfers to a previously unseen voice-personalization backend.

Table 21 shows that LEAP transfers across all three held-out voice-personalization models. It exceeds learned static ranking by $+0.022$ on held-out Seed-VC, $+0.023$ on OpenVoice V2, and $+0.022$ on FreeVC, despite receiving no utility outcomes from the respective target model during policy learning or model selection.

This result is stronger than showing that a single fixed enrollment script is robust across backends. Learned static ranking is trained on exactly the same source-model outcomes and uses the same text-side linguistic descriptors as LEAP, yet it consistently underperforms the dynamic, state-aware

Table 14: **Per-language preference in the preregistered primary comparison.** Values are speaker-similarity judgments favoring LEAP over learned static ranking, pooled over 3- and 7-second budgets. Each language contributes 144 retained judgments.

Language	LEAP preference	95% CI	n
Korean	59.0%	[50.6, 67.0]	144
Japanese	60.4%	[52.0, 68.3]	144
Vietnamese	59.7%	[51.2, 67.7]	144
Pooled	59.7%	[55.1, 64.1]	432

Table 15: **Confirmatory mixed-effects result for the preregistered primary human comparison.** The coefficient is on the log-odds scale for selecting LEAP over learned static ranking.

Comparison	β	Odds ratio	95% CI (OR)	p
LEAP vs static, speaker similarity	0.401	1.49	[1.18, 1.89]	0.0008

policy. Thus, LEAP transfers not only a global preference over prompts, but also a model-robust strategy for selecting complementary enrollment evidence under a recording-time constraint.

LEAP also outperforms the language-specific handcrafted objective for each held-out model. Handcrafted coverage remains useful across architectures, but it cannot adapt the value of a candidate to evidence already represented in the current enrollment set. Together, these results indicate that the offline utility learner captures selection regularities shared by Seed-VC, OpenVoice V2, and FreeVC, allowing the policy to be applied to a voice-personalization backend unseen during utility learning without collecting target-model utility labels. Per-budget, component-wise, and statistical results are provided in Appendix C.7.

C.7 Detailed Results for Outcome-Free Model Transfer

Outcome-free model calibration. For a leave-one-model-out target model $m^\star$ and automatic metric k, we compute calibration parameters

$$\mu_k^{-m^\star},\ \sigma_k^{-m^\star}$$

from pooled training-speaker outcomes of the two source voice-personalization models, across the source languages only. We freeze these parameters before evaluation and apply

$$\widetilde{M}_k^{-m^\star}(S; s, m^\star) = \frac{M_k(S; s, m^\star) - \mu_k^{-m^\star}}{\sigma_k^{-m^\star} + \epsilon}$$

unchanged to the held-out target model. This source-model-only evaluation calibration is distinct from the source-condition target standardization used during offline utility learning (Appendix A.3). Thus, the held-out model contributes no utility-learning outcomes, calibration statistics, validation scores, checkpoint-selection signals, or policy updates.

Statistical analysis. For each held-out target model and method, we first average utility within each test speaker across languages, recording budgets, and evaluation utterances. We then perform paired speaker-level bootstrap comparisons between LEAP and learned static ranking using 10,000 resamples. Holm–Bonferroni correction is applied over the three held-out target-model comparisons.

C.8 What Linguistic Evidence Does LEAP Select?

The performance gains of LEAP raise a complementary question: what evidence does the policy select under a limited recording budget? We analyze the linguistic annotations of enrollment sets generated in held-out evaluation folds, focusing on the 3- and 5-second conditions where content selection is most consequential. To compare languages with different phonological systems, we map language-specific annotations to the shared functional families defined in Appendix A.1: segmental units;

Table 16: **Robustness to action collection.** Utility (↑) is averaged over Korean, Japanese, Vietnamese, and downstream voice-personalization models. Values are mean ± standard deviation over three policy-training seeds; all methods use identical speaker-disjoint partitions and test-time selection over all feasible candidates.

Offline action collection	**3 s**	**Avg.**
Default: 8 uniform + 4 coverage + 4 exploratory	**0.652 ± 0.006**	**0.694 ± 0.004**
32 actions: 16 uniform + 8 coverage + 8 exploratory	0.650 ± 0.007	0.692 ± 0.005
All feasible actions	0.651 ± 0.006	0.693 ± 0.004
Uniform only: 16 uniform	0.638 ± 0.008	0.682 ± 0.006

Table 17: **Generalization beyond utility-learning conditions.** Values are normalized composite personalization utility (↑), averaged over Korean, Japanese, and Vietnamese. All settings use speaker-disjoint test speakers. For unseen-budget evaluation, outcomes from the target budget are excluded during policy training and model selection. For unseen lexical realizations, test prompts are alternative instantiations of the same linguistic category specifications and are absent from the utility-learning prompt inventory. † denotes a significant improvement over learned static ranking under a paired speaker-level test ($p < 0.05$, Holm corrected).

Axis	**Test condition**	**Static**	**LEAP**	**Δ**
Unseen speakers	3 s	0.619	**0.652**†	+0.033
	10 s	0.710	**0.727**†	+0.017
Unseen budgets	5 s held out during training	0.655	**0.681**†	+0.026
	10 s held out during training	0.706	**0.722**†	+0.016
Unseen lexical realizations	3 s	0.612	**0.641**†	+0.029
	10 s	0.704	**0.723**†	+0.019

syllabic, moraic, and coda structure; prosodic contrasts and pronunciation processes; within-utterance transitions; and cross-boundary and visemic transition structure.

For each selected enrollment set, we compute coverage as the proportion of canonical descriptor types in each family represented by the union of its utterances, normalized by the corresponding language-specific candidate inventory. For a selection trajectory $x_1, \ldots, x_T$, we additionally measure marginal evidence overlap as the mean, across selected utterances and applicable families, of the proportion of descriptors in x_t that are already represented by $\{x_1, \ldots, x_{t-1}\}$. These statistics are computed solely from text-side annotations and selected trajectories; downstream personalization outcomes are not used to calculate the reported coverage or overlap measures.

Table 25 shows that LEAP does not merely maximize phoneme coverage. The phoneme-coverage baseline represents the largest proportion of segmental units, but it underrepresents prosodic, process-related, and transition evidence. In contrast, LEAP achieves the highest coverage for every non-segmental evidence family. Its largest advantages over learned static ranking occur for prosodic contrasts and pronunciation processes ($+13.3$ points) and for cross-boundary and visemic transitions ($+11.7$ points). LEAP also selects less redundant evidence over the course of an enrollment trajectory: its mean marginal overlap is 0.29, compared with 0.46 for learned static ranking.

Table 26 further summarizes the aggregate set-level properties by budget. LEAP selects sets with higher canonical-descriptor coverage, more unique descriptors, and substantially lower within-set redundancy than learned static ranking at both short budgets. Providing remaining-budget information to the static ranker yields only a small improvement, whereas removing the remaining-budget input from LEAP reduces performance (Table 27). Thus, budget awareness alone does not account for the gain of the full policy; re-scoring candidates relative to the evolving enrollment set is also important for constructing complementary content. The Korean trajectory in Figure 1 provides a concrete example of this behavior.

Table 18: **Per-language generalization results.** Normalized composite personalization utility (↑). Each cell reports learned static ranking / LEAP; the macro-average over the three languages reproduces Table 17.

Axis	Test condition	Korean	Japanese	Vietnamese	Macro-average
Unseen speakers	3 s	0.625 / **0.659**	0.614 / **0.644**	0.618 / **0.653**	0.619 / **0.652**
	10 s	0.716 / **0.734**	0.706 / **0.720**	0.708 / **0.727**	0.710 / **0.727**
Unseen budgets	5 s held out	0.659 / **0.686**	0.651 / **0.677**	0.655 / **0.680**	0.655 / **0.681**
	10 s held out	0.711 / **0.727**	0.702 / **0.718**	0.705 / **0.721**	0.706 / **0.722**
Unseen lexical realizations	3 s	0.618 / **0.648**	0.607 / **0.635**	0.611 / **0.640**	0.612 / **0.641**
	10 s	0.710 / **0.729**	0.700 / **0.719**	0.702 / **0.721**	0.704 / **0.723**

Table 19: **Generalization under resampled offline action collections.** Mean ± standard deviation of normalized composite utility over independent action-collection resampling seeds.

Axis	Condition	Static	LEAP
Unseen speakers	3 s	0.619 ± 0.004	$\mathbf{0.652 \pm 0.005}$
Unseen speakers	10 s	0.710 ± 0.003	$\mathbf{0.727 \pm 0.004}$
Unseen budgets	5 s held out	0.655 ± 0.004	$\mathbf{0.681 \pm 0.005}$
Unseen budgets	10 s held out	0.706 ± 0.003	$\mathbf{0.722 \pm 0.004}$
Unseen lexical realizations	3 s	0.612 ± 0.005	$\mathbf{0.641 \pm 0.006}$
Unseen lexical realizations	10 s	0.704 ± 0.004	$\mathbf{0.723 \pm 0.005}$

Table 20: **Component-wise generalization results.** Metrics are averaged over the two conditions reported for each generalization axis. Composite utility is the unweighted mean of the three normalized components.

Axis	Method	Speaker sim.	Content pres.	Naturalness	Composite
Unseen speakers	Static	0.681	0.662	0.651	0.665
	LEAP	**0.706**	**0.692**	**0.671**	**0.690**
Unseen budgets	Static	0.694	0.680	0.668	0.681
	LEAP	**0.715**	**0.704**	**0.686**	**0.702**
Unseen lexical realizations	Static	0.674	0.656	0.644	0.658
	LEAP	**0.697**	**0.680**	**0.669**	**0.682**

Table 21: **Transfer to unseen voice-personalization models.** Values are normalized composite personalization utility (↑), averaged over Korean, Japanese, Vietnamese, and all recording budgets. For each row, the target model is excluded from offline utility learning; LEAP and learned static ranking use outcomes from the other two models only. For each leave-one-model-out fold, metric calibration parameters are estimated from training-speaker outcomes of the two source models only, frozen before evaluation, and applied unchanged to the held-out target model. † denotes a significant improvement over learned static ranking under a paired speaker-level test ($p < 0.05$, Holm corrected).

Held-out target model	Phoneme cov.	Handcraft	Static	LEAP
Seed-VC	0.641	0.657	0.664	**0.686**†
OpenVoice V2	0.628	0.646	0.651	**0.674**†
FreeVC	0.617	0.635	0.641	**0.663**†
Average	0.629	0.646	0.652	**0.674**

Table 22: **Per-budget transfer to unseen voice-personalization models.** Normalized composite personalization utility (↑). Each target model is excluded from offline utility learning; the average reproduces Table 21.

Held-out model	Method	3 s	5 s	7 s	10 s	Avg.
Seed-VC	Phoneme coverage	0.604	0.634	0.656	0.670	0.641
	Language handcrafted	0.621	0.650	0.671	0.686	0.657
	Learned static ranking	0.625	0.656	0.680	0.695	0.664
	LEAP	**0.648**	**0.680**	**0.705**	**0.711**	**0.686**
OpenVoice V2	Phoneme coverage	0.571	0.612	0.648	0.681	0.628
	Language handcrafted	0.591	0.631	0.665	0.697	0.646
	Learned static ranking	0.611	0.644	0.671	0.678	0.651
	LEAP	**0.635**	**0.669**	**0.693**	**0.699**	**0.674**
FreeVC	Phoneme coverage	0.561	0.604	0.641	0.662	0.617
	Language handcrafted	0.580	0.623	0.658	0.679	0.635
	Learned static ranking	0.601	0.632	0.657	0.674	0.641
	LEAP	**0.624**	**0.657**	**0.680**	**0.691**	**0.663**

Table 23: **Component-wise unseen-model transfer results.** Metrics are source-model-normalized and averaged over Korean, Japanese, Vietnamese, and the four recording budgets.

Held-out model	Method	Speaker sim.	Content pres.	Naturalness	Composite
Seed-VC	Static	0.679	0.664	0.649	0.664
	LEAP	**0.702**	**0.689**	**0.667**	**0.686**
OpenVoice V2	Static	0.665	0.651	0.637	0.651
	LEAP	**0.688**	**0.677**	**0.657**	**0.674**
FreeVC	Static	0.654	0.641	0.628	0.641
	LEAP	**0.676**	**0.665**	**0.648**	**0.663**

Table 24: **Statistical comparison for unseen-model transfer.** Paired speaker-level bootstrap comparisons test LEAP against learned static ranking. Confidence intervals are for composite-utility differences.

Held-out model	Δ utility	95% CI	Effect size	p_{Holm}
Seed-VC	+0.022	[0.014, 0.030]	0.56	0.002
OpenVoice V2	+0.023	[0.015, 0.031]	0.60	< 0.001
FreeVC	+0.022	[0.013, 0.031]	0.53	0.003

Table 25: **Linguistic evidence represented by selected enrollment sets.** Coverage values are percentages, computed from the union of text-side linguistic annotations in the final enrollment set and normalized within each language-specific candidate inventory. Results are averaged over Korean, Japanese, Vietnamese, and the 3- and 5-second budgets. Higher values indicate that more distinct evidence of the corresponding family is represented. The final row reports marginal evidence overlap between a newly selected utterance and evidence already present in the enrollment set (↓).

Evidence family	Phoneme-coverage baseline	Learned static	LEAP
Segmental units	**64.7**	58.4	60.2
Syllabic, moraic, and coda structure	59.1	65.1	**72.4**
Prosodic contrasts and pronunciation processes	36.8	47.5	**60.8**
Within-utterance transitions	12.2	15.6	**23.9**
Cross-boundary and visemic transitions	18.1	24.0	**35.7**
Marginal evidence overlap (↓)	0.39	0.46	**0.29**

Table 26: **Set-level evidence properties of selected enrollment sets.** Canonical-descriptor coverage is normalized by the language-specific candidate inventory. Pairwise redundancy is the mean Jaccard overlap between canonical descriptor sets of selected utterances; lower is better. Results are macro-averaged across primary languages and held-out evaluation folds.

Budget	Method	Coverage (↑)	Unique desc. (↑)	Redundancy (↓)
3 s	Learned static ranking	43.1%	34.7	0.374
	LEAP	**51.8%**	**41.6**	**0.219**
5 s	Learned static ranking	58.4%	49.1	0.331
	LEAP	**65.7%**	**55.8**	**0.201**

Table 27: **Ablating state and remaining-budget information.** All variants use the same text-side linguistic annotations and offline utility supervision. "Static + budget" receives remaining-budget features but no representation of the selected set.

Method	3 s	5 s	7 s	10 s	Avg.
Learned static ranking	0.619	0.659	0.686	0.710	0.669
Static + remaining budget	0.624	0.663	0.690	0.713	0.673
LEAP w/o remaining-budget input	0.645	0.681	0.704	0.723	0.688
LEAP	**0.652**	**0.687**	**0.709**	**0.727**	**0.694**

D Full Lexicon and Additional Enrollment Sequences

Tables 28–31 provide the complete enrollment lexicons for Korean, Japanese, Vietnamese, and English, including durations, canonical pronunciations, linguistic annotations, component-wise coverage gains, and utility scores.

Table 28: Full lexicon table ($N = 96$), consisting of words, phrases, and sentences. The table reports mean actual duration (Dur), canonical pronunciation (Pron), coda inventory (Co), phoneme transitions (Tr), morphophonological processes (Mo), phoneme coverage gain, coda coverage gain, morphophonological coverage gain, phoneme-transition coverage gain (ptr), and audio score per second.

ID	Type	Text	Dur.	Pron.	Co	Tr	Mo	C_{ph}	C_{coda}	C_{morph}	C_{ptr}	U_A
K001	word	바다	0.63	바다 (bada)		ㅂ → ㅏ, ㅏ → ㄷ ㄷ → ㅏ		0.06	0	0	0.011	0.03
K002	word	가방	0.6	가방 (gabang)	ㅇ	ㄱ → ㅏ, ㅏ → ㅂ ㅂ → ㅏ, ㅏ → ㅇ		0.08	0.09	0	0.01	0.08
K003	word	다리	0.53	다리 (dari)		ㄷ → ㅏ, ㅏ → ㄹ ㄹ → ㅣ		0.08	0	0	0.01	0.04
K004	word	사과	0.63	사과 (sagwa)		ㅅ → ㅏ, ㅏ → ㄱ ㄱ → ㅘ		0.08	0	0	0.01	0.04
K005	word	자두	0.47	자두 (jadu)		ㅈ → ㅏ, ㅏ → ㄷ ㄷ → ㅜ		0.08	0	0	0.011	0.05
K006	word	까치	0.57	까치 (kkachi)		ㄲ → ㅏ, ㅏ → ㅊ ㅊ → ㅣ		0.08	0	0	0.011	0.04
K007	word	딸기	0.77	딸기 (ttalgi)	ㄹ	ㄸ → ㅏ, ㅏ → ㄹ ㄹ → ㄱ, ㄱ → ㅣ		0.1	0.09	0	0.01	0.07
K008	word	빨대	0.8	빨때 (ppalttae)	ㄹ	ㅃ → ㅏ, ㅏ → ㄹ ㄹ → ㄸ, ㄸ → ㅐ	tensification	0.1	0.09	0.13	0.01	0.10
K009	word	쌀밥	0.77	쌀밥 (ssalbap)	ㄹ, ㅂ	ㅆ → ㅏ, ㅏ → ㄹ ㄹ → ㅂ, ㅂ → ㅏ ㅏ → ㅂ		0.08	0.18	0	0.02	0.09
K010	word	짜장	0.6	짜장 (jjajang)	ㅇ	ㅉ → ㅏ, ㅏ → ㅈ ㅈ → ㅏ, ㅏ → ㅇ		0.08	0.09	0	0.01	0.08
K011	word	카드	0.63	카드 (kadeu)		ㅋ → ㅏ, ㅏ → ㄷ ㄷ → ㅡ		0.08	0	0	0.01	0.04
K012	word	토끼	0.6	토끼 (tokki)		ㅌ → ㅗ, ㅗ → ㄲ ㄲ → ㅣ		0.08	0	0	0.01	0.04
K013	word	포도	0.67	포도 (podo)		ㅍ → ㅗ, ㅗ → ㄷ ㄷ → ㅗ		0.06	0	0	0.01	0.03
K014	word	차표	0.77	차표 (chapyo)		ㅊ → ㅏ, ㅏ → ㅍ ㅍ → ㅛ		0.08	0	0	0.011	0.03
K015	word	커피	0.7	커피 (keopi)		ㅋ → ㅓ, ㅓ → ㅍ ㅍ → ㅣ		0.08	0	0	0.011	0.03
K016	word	나무	0.5	나무 (namu)		ㄴ → ㅏ, ㅏ → ㅁ ㅁ → ㅜ		0.08	0	0	0.01	0.06
K017	word	마음	0.5	마음 (maeum)	ㅁ	ㅁ → ㅏ, ㅏ → ㅇ ㅇ → ㅡ, ㅡ → ㅁ		0.08	0.09	0	0.01	0.09
K018	word	안녕	0.6	안녕 (annyeong)	ㄴ, ㅇ	ㅇ → ㅏ, ㅏ → ㄴ ㄴ → ㄴ, ㄴ → ㅕ ㅕ → ㅇ		0.08	0.18	0	0.01	0.12
K019	word	강물	0.7	강물 (gangmul)	ㅇ, ㄹ	ㄱ → ㅏ, ㅏ → ㅇ ㅇ → ㅁ, ㅁ → ㅜ ㅜ → ㄹ		0.12	0.18	0	0.02	0.11
K020	word	라면	0.57	라면 (ramyeon)	ㄴ	ㄹ → ㅏ, ㅏ → ㅁ ㅁ → ㅕ, ㅕ → ㄴ		0.1	0.09	0	0.01	0.09
K021	word	노래	0.53	노래 (norae)		ㄴ → ㅗ, ㅗ → ㄹ ㄹ → ㅐ		0.08	0	0	0.01	0.04
K022	word	거리	0.57	거리 (geori)		ㄱ → ㅓ, ㅓ → ㄹ ㄹ → ㅣ		0.08	0	0	0.01	0.04
K023	word	국물	0.6	궁물 (gungmul)	ㅇ, ㄹ	ㄱ → ㅜ, ㅜ → ㅇ ㅇ → ㅁ, ㅁ → ㅜ ㅜ → ㄹ	nasal assimilation	0.1	0.18	0.13	0.02	0.18
K024	word	신라	0.6	실라 (silla)	ㄹ	ㅅ → ㅣ, ㅣ → ㄹ ㄹ → ㄹ, ㄹ → ㅏ	liquid assimilation	0.08	0.09	0.12	0.01	0.13
K025	word	같이	0.53	가치 (gachi)		ㄱ → ㅏ, ㅏ → ㅊ ㅊ → ㅇ, ㅇ → ㅣ	palatalization	0.08	0	0.13	0.01	0.10
K026	word	학교	0.63	학꾜 (hakkkyo)	ㄱ	ㅎ → ㅏ, ㅏ → ㄱ ㄱ → ㄲ, ㄲ → ㅛ	tensification	0.08	0.09	0.13	0.01	0.12
K027	word	좋다	0.73	조타 (jota)		ㅈ → ㅗ, ㅗ → ㅌ ㅌ → ㅏ	aspiration	0.08	0	0.13	0.01	0.07

Continued on next page

K028	word	옷이	0.6	오시 (osi)		ㅇ→ㅗ, ㅗ→ㅅ ㅅ→ ㅣ	resyllabification	0.08	0	0.13	0.01	0.09
K029	word	앞문	0.7	암문 (ammun)	ㅁ, ㄴ	ㅇ→ ㅏ, ㅏ→ㅁ ㅁ→ㅁ, ㅁ→ㅜ ㅜ→ㄴ	nasal assimilation	0.1	0.18	0.13	0.02	0.15
K030	word	꽃말	0.67	꼰말 (kkonmal)	ㄴ, ㄹ	ㄲ→ㄱ, ㄱ→ㄴ ㄴ→ㅁ, ㅁ→ ㅏ ㅏ→ㄹ	nasal assimilation	0.12	0.18	0.13	0.02	0.17
K031	word	먹는	0.63	멍는 (meongneun)	ㅇ,ㄴ	ㅁ→ ㅓ, ㅓ→ㅇ ㅇ→ㄴ, ㄴ→ㅡ ㅡ→ㄴ	nasal assimilation	0.1	0.18	0.13	0.02	0.17
K032	word	닫는	0.53	단는 (danneun)	ㄴ	ㄷ→ ㅏ, ㅏ→ㄴ ㄴ→ㄴ, ㄴ→ㅡ ㅡ→ㄴ	nasal assimilation	0.08	0.09	0.13	0.02	0.15
K033	word	설날	0.57	설랄 (seollal)	ㄹ	ㅅ→ ㅓ, ㅓ→ㄹ ㄹ→ㄹ, ㄹ→ ㅏ ㅏ→ㄹ	liquid assimilation	0.08	0.09	0.13	0.02	0.14
K034	word	색연필	0.93	생년필 (saeng nyeonpil)	ㅇ, ㄴ, ㄹ	ㅅ→ ㅐ, ㅐ→ㅇ ㅇ→ㄴ, ㄴ→ㅕ ㅕ→ㄴ, ㄴ→ㅍ ㅍ→ ㅣ, ㅣ→ㄹ	nasal assimilation	0.16	0.27	0.13	0.03	0.16
K035	word	독립	0.67	동닙 (dongnip)	ㅇ, ㅂ	ㄷ→ㅗ, ㅗ→ㅇ ㅇ→ㄴ, ㄴ→ ㅣ ㅣ→ㅂ	nasal assimilation	0.12	0.18	0.13	0.02	0.17
K036	word	아이	0.47	아이 (ai)		ㅇ→ ㅏ, ㅏ→ㅇ ㅇ→ ㅣ		0.06	0	0	0.011	0.04
K037	word	오이	0.5	오이 (oi)		ㅇ→ㅗ, ㅗ→ㅇ ㅇ→ ㅣ		0.06	0	0	0.01	0.04
K038	word	우유	0.5	우유 (uyu)		ㅇ→ㅜ, ㅜ→ㅇ ㅇ→ㅠ		0.06	0	0	0.01	0.04
K039	word	여우	0.5	여우 (yeou)		ㅇ→ ㅕ, ㅕ→ㅇ ㅇ→ㅜ		0.06	0	0	0.01	0.04
K040	word	이유	0.4	이유 (iyu)		ㅇ→ ㅣ, ㅣ→ㅇ ㅇ→ㅠ		0.06	0	0	0.01	0.04
K041	word	바구니	0.8	바구니 (baguni)		ㅂ→ ㅏ, ㅏ→ㄱ ㄱ→ㅜ, ㅜ→ㄴ ㄴ→ ㅣ		0.12	0	0	0.02	0.04
K042	word	고구마	0.73	고구마 (goguma)		ㄱ→ㅗ, ㅗ→ㄱ ㄱ→ㅜ, ㅜ→ㅁ ㅁ→ ㅏ		0.1	0	0	0.02	0.04
K043	word	비빔밥	0.97	비빔빱 (bibimppap)	ㅁ,ㅂ	ㅂ→ ㅣ, ㅣ→ㅂ ㅂ→ ㅣ, ㅣ→ㅁ ㅁ→ㅃ, ㅃ→ ㅏ ㅏ→ㅂ	tensification	0.08	0.18	0.13	0.02	0.11
K044	word	보리밥	0.83	보리밥 (boribap)	ㅂ	ㅂ→ㅗ, ㅗ→ㄹ ㄹ→ ㅣ, ㅣ→ㅂ ㅂ→ ㅏ, ㅏ→ㅂ		0.1	0.09	0	0.02	0.06
K045	word	카메라	0.77	카메라 (kamera)		ㅋ→ ㅏ, ㅏ→ㅁ ㅁ→ ㅔ, ㅔ→ㄹ ㄹ→ ㅏ		0.1	0	0	0.02	0.04
K046	word	토마토	0.73	토마토 (tomato)		ㅌ→ㅗ, ㅗ→ㅁ ㅁ→ ㅏ, ㅏ→ㅌ ㅌ→ㅗ		0.08	0	0	0.01	0.03
K047	word	운동화	0.8	운동화 (undonghwa)	ㄴ, ㅇ	ㅇ→ㅜ, ㅜ→ㄴ ㄴ→ㄷ, ㄷ→ㅗ ㅗ→ㅇ, ㅇ→ㅎ ㅎ→ㅘ		0.14	0.18	0	0.03	0.12
K048	word	원피스	0.77	원피스 (wonpiseu)	ㄴ	ㅇ→ㅝ, ㅝ→ㄴ ㄴ→ㅍ, ㅍ→ ㅣ ㅣ→ㅅ, ㅅ→ㅡ		0.14	0.09	0	0.02	0.08
K049	word	초콜릿	0.8	초콜릳 (chokollit)	ㄹ,ㄷ	ㅊ→ㅗ, ㅗ→ㅋ ㅋ→ㅗ, ㅗ→ㄹ ㄹ→ㄹ, ㄹ→ ㅣ ㅣ→ㄷ	coda neutralization	0.12	0.18	0.13	0.03	0.14
K050	word	빨간색	1.1	빨간색 (ppalgansaek)	ㄹ,ㄴ,ㄱ	ㅃ→ ㅏ, ㅏ→ㄹ ㄹ→ㄱ, ㄱ→ ㅏ ㅏ→ㄴ, ㄴ→ㅅ ㅅ→ ㅐ, ㅐ→ㄱ		0.14	0.27	0	0.03	0.10
K051	word	텔레비전	0.93	텔레비전 (tellebijeon)	ㄹ,ㄴ	ㅌ→ ㅔ, ㅔ→ㄹ ㄹ→ㄹ, ㄹ→ ㅖ ㅖ→ㅂ, ㅂ→ ㅣ ㅣ→ㅈ, ㅈ→ㅓ ㅓ→ㄴ		0.16	0.18	0	0.03	0.10
K052	word	아버지	0.77	아버지 (abeoji)		ㅇ→ ㅏ, ㅏ→ㅂ ㅂ→ ㅓ, ㅓ→ㅈ ㅈ→ ㅣ		0.12	0	0	0.02	0.04
K053	word	어머니	0.77	어머니 (eomeoni)		ㅇ→ ㅓ, ㅓ→ㅁ ㅁ→ ㅓ, ㅓ→ㄴ ㄴ→ ㅣ		0.1	0	0	0.02	0.04

Continued on next page

K054	word	엘리베이터	1.07	엘리베이터 (ellibeiteo)	ㄹ	ㅇ → ㅖ, ㅖ → ㄹ ㄹ → ㄹ, ㄹ → ㅣ ㅣ → ㅂ, ㅂ → ㅖ ㅖ → ㅇ, ㅇ → ㅣ ㅣ → ㅌ, ㅌ → ㅓ		0.14	0.09	0	0.04	0.06
K055	word	아이스크림	1	아이스크림 (aiseukeurim)	ㅁ	ㅇ → ㅏ, ㅏ → ㅇ ㅇ → ㅣ, ㅣ → ㅅ ㅅ → ㅡ, ㅡ → ㅋ ㅋ → ㅡ, ㅡ → ㄹ ㄹ → ㅣ, ㅣ → ㅁ		0.16	0.09	0	0.04	0.07
K056	word	아메리카노	1.13	아메리카노 (amerikano)		ㅇ → ㅏ, ㅏ → ㅁ ㅁ → ㅖ, ㅖ → ㄹ ㄹ → ㅣ, ㅣ → ㅋ ㅋ → ㅏ, ㅏ → ㄴ ㄴ → ㄴ		0.18	0	0	0.03	0.05
K057	phrase	밥 먹다	1.13	밤 먹따 (bam meoktta)	ㅁ, ㄱ	ㅂ → ㅏ, ㅏ → ㅁ ㅁ → ㅁ, ㅁ → ㅓ ㅓ → ㄱ, ㄱ → ㄲ ㄲ → ㅏ	nasal assimilation tensification	0.12	0.18	0.25	0.03	0.13
K058	phrase	물 마시다	1.1	물 마시다 (mul masida)	ㄹ	ㅁ → ㅜ, ㅜ → ㄹ ㄹ → ㅁ ㅁ → ㅏ, ㅏ → ㅅ ㅅ → ㅣ, ㅣ → ㄷ ㄷ → ㅏ		0.14	0.09	0	0.03	0.06
K059	phrase	문 열다	0.97	문 녈다 (mun nyeolda)	ㄴ, ㄹ	ㅁ → ㅜ, ㅜ → ㄴ ㄴ → ㄴ ㄴ → ㅕ, ㅕ → ㄹ ㄹ → ㄷ, ㄷ → ㅏ	n insertion	0.14	0.18	0.13	0.03	0.12
K060	phrase	문 닫다	0.97	문 닫따 (mun dattta)	ㄴ, ㄷ	ㅁ → ㅜ, ㅜ → ㄴ ㄴ → ㄷ ㄷ → ㅏ, ㅏ → ㄷ ㄷ → ㄸ, ㄸ → ㅏ	tensification	0.12	0.18	0.13	0.03	0.12
K061	phrase	옷 입다	1.17	온 닙따 (onniptta)	ㄴ, ㅂ	ㅇ → ㅗ, ㅗ → ㄴ ㄴ → ㅇ ㅇ → ㅣ, ㅣ → ㅂ ㅂ → ㅃ, ㅃ → ㅏ	n insertion, nasal assimilation tensification	0.14	0.18	0.38	0.03	0.16
K062	phrase	책 읽다	1.13	챙 닉따 (chaeng nikkta)	ㅇ, ㄱ	ㅊ → ㅐ, ㅐ → ㅇ ㅇ → ㄴ ㄴ → ㅣ, ㅣ → ㄱ ㄱ → ㄲ, ㄲ → ㅏ	n insertion	0.16	0.18	0.13	0.03	0.11
K063	phrase	커피 마시다	1.1	커피 마시다 (keopi masida)		ㅋ → ㅓ, ㅓ → ㅍ ㅍ → ㅣ, ㅣ → ㅁ ㅁ → ㅏ, ㅏ → ㅅ ㅅ → ㅣ, ㅣ → ㄷ ㄷ → ㅏ		0.16	0	0	0.04	0.04
K064	phrase	우유 마시다	1.27	우유 마시다 (uyu masida)		ㅇ → ㅜ, ㅜ → ㅇ ㅇ → ㅠ ㅠ → ㅁ, ㅁ → ㅏ ㅏ → ㅅ, ㅅ → ㅣ ㅣ → ㄷ, ㄷ → ㅏ		0.16	0	0	0.04	0.04
K065	phrase	밥 먹어요	1.23	밤 머거요 (bam meogeoyo)	ㅁ	ㅂ → ㅏ, ㅏ → ㅁ ㅁ → ㅁ ㅁ → ㅓ, ㅓ → ㄱ ㄱ → ㅇ, ㅇ → ㅓ ㅓ → ㅇ, ㅇ → ㅛ	nasal assimilation resyllabification	0.14	0.09	0.25	0.04	0.11
K066	phrase	물 주세요	1.1	물 주세요 (mul juseyo)	ㄹ	ㅁ → ㅜ, ㅜ → ㄹ ㄹ → ㅈ ㅈ → ㅜ, ㅜ → ㅅ ㅅ → ㅖ, ㅖ → ㅇ ㅇ → ㅛ		0.16	0.09	0	0.03	0.06
K067	phrase	문 녀러	1.07	문 녀러 (mun nyeoreo)	ㄴ	ㅁ → ㅜ, ㅜ → ㄴ ㄴ → ㄴ ㄴ → ㅕ, ㅕ → ㄹ ㄹ → ㅓ	n insertion, resyllabification	0.12	0.09	0.25	0.03	0.11
K068	phrase	문 닫아	1.2	문 다다 (mun dada)	ㄴ	ㅁ → ㅜ, ㅜ → ㄴ ㄴ → ㄷ ㄷ → ㅏ, ㅏ → ㄷ ㄷ → ㄷ, ㄷ → ㅏ	resyllabification	0.1	0.09	0.13	0.02	0.07
K069	phrase	빨리 와	1.03	빨리 와 (ppalli wa)	ㄹ	ㅃ → ㅏ, ㅏ → ㄹ ㄹ → ㄹ, ㄹ → ㅣ ㅣ → ㅇ ㅇ → ㅘ		0.12	0.09	0	0.03	0.06
K070	phrase	같이 가요	1	가치 가요 (gachi gayo)		ㄱ → ㅏ, ㅏ → ㅊ ㅊ → ㅣ, ㄱ → ㅏ ㅏ → ㅇ, ㅇ → ㅛ	palatalization	0.12	0	0.13	0.02	0.07
K071	phrase	집에 가다	0.93	지베 가다 (jibe gada)		ㅈ → ㅣ, ㅣ → ㅂ ㅂ → ㅖ, ㅖ → ㄱ ㄱ → ㅏ ㅏ → ㄷ, ㄷ → ㅏ	resyllabification	0.14	0	0.13	0.03	0.09

Continued on next page

K072	phrase	학교 가다	0.83	학꾜 가다 (hakkkyo gada)	ㄱ	ㅎ→ㅏ, ㅏ→ㄱ ㄱ→ㄲ, →ㅛ ㅛ→ㄱ ㄱ→ㅏ, ㅏ→ㄷ ㄷ→ㅏ	tensification	0.12	0.09	0.13	0.03	0.11
K073	phrase	사진 찍다	1.13	사진 찍따 (sajin jjiktta)	ㄴ, ㄱ	ㅅ→ㅏ, ㅏ→ㅈ ㅈ→ㅣ, ㅣ→ㄴ ㄴ→ㅉ ㅉ→ㅣ, ㅣ→ㄱ ㄱ→ㄷ, ㄷ→ㅏ	tensification	0.16	0.18	0.13	0.04	0.11
K074	phrase	전화 받다	1.1	전화 받따 (jeonhwa battta)	ㄴ, ㄷ	ㅈ→ㅓ, ㅓ→ㄴ ㄴ→ㅎ, ㅎ→ㅘ ㅘ→ㅂ ㅂ→ㅏ, ㅏ→ㄷ ㄷ→ㄸ, ㄸ→ㅏ	tensification	0.18	0.18	0.13	0.04	0.12
K075	phrase	음악 듣다	1.2	으막 듣따 (eumak deuttta)	ㄱ, ㄷ	ㅇ→ㅡ, ㅡ→ㅁ ㅁ→ㅁ, ㅁ→ㅏ ㅏ→ㄱ, ㄱ→ㄷ ㄷ→ㅡ ㅡ→ㄷ, ㄷ→ㄸ ㄸ→ㅏ	resyllabification, tensification	0.14	0.18	0.25	0.04	0.13
K076	phrase	이름 부르다	1.23	이름 부르다 (ireum bureuda)	ㅁ	ㅇ→ㅣ, ㅣ→ㄹ ㄹ→ㅡ, ㅡ→ㅁ ㅁ→ㅂ ㅂ→ㅜ, ㅜ→ㄹ ㄹ→ㅡ, ㅡ→ㄷ ㄷ→ㅏ		0.2	0.09	0	0.04	0.07
K077	phrase	얼굴 보다	1.07	얼굴 보다 (eolgul boda)	ㄹ	ㅇ→ㅓ, ㅓ→ㄹ ㄹ→ㄱ, ㄱ→ㅜ ㅜ→ㄹ, ㄹ→ㅂ ㅂ→ㅗ ㅗ→ㄷ, ㄷ→ㅏ		0.18	0.09	0	0.04	0.07
K078	phrase	친구 만나다	1.13	친구 만나다 (chingu mannada)	ㄴ	ㅊ→ㅣ, ㅣ→ㄴ ㄴ→ㄱ, ㄱ→ㅜ ㅜ→ㅁ ㅁ→ㅏ, ㅏ→ㄴ ㄴ→ㄴ, ㄴ→ㅏ ㅏ→ㄷ, ㄷ→ㅏ		0.16	0.09	0	0.04	0.07
K079	phrase	버스 타다	1.03	버스 타다 (beoseu tada)		ㅂ→ㅓ, ㅓ→ㅅ ㅅ→ㅡ, ㅡ→ㅌ ㅌ→ㅏ ㅏ→ㄷ, ㄷ→ㅏ		0.14	0	0	0.03	0.04
K080	phrase	택시 타다	1.07	택씨 타다 (taekssi tada)	ㄱ	ㅌ→ㅐ, ㅐ→ㄱ ㄱ→ㅆ, ㅆ→ㅣ ㅣ→ㅌ ㅌ→ㅏ, ㅏ→ㄷ ㄷ→ㅏ	tensification	0.14	0.09	0.13	0.03	0.09
K081	phrase	집에 오다	0.97	지베 오다 (jibe oda)		ㅈ→ㅣ, ㅣ→ㅂ ㅂ→ㅖ, ㅖ→ㅇ ㅇ→ㅗ ㅗ→ㄷ, ㄷ→ㅏ	resyllabification	0.14	0	0.13	0.03	0.08
K082	phrase	밥을 먹다	1.03	바블 먹따 (babeul meoktta)	ㄹ, ㄱ	ㅂ→ㅏ, ㅏ→ㅂ ㅂ→ㅡ, ㅡ→ㄹ ㄹ→ㅁ ㅁ→ㅓ, ㅓ→ㄱ ㄱ→ㄸ, ㄸ→ㅏ	resyllabification, tensification	0.18	0.18	0.25	0.04	0.16
K083	phrase	물을 먹다	1	무를 먹따 (mureul meoktta)	ㄹ, ㄱ	ㅁ→ㅜ, ㅜ→ㄹ ㄹ→ㅡ, ㅡ→ㄹ ㄹ→ㅁ, ㅁ→ㅓ ㅓ→ㄱ, ㄱ→ㄸ ㄸ→ㅏ	resyllabification, tensification	0.16	0.18	0.25	0.04	0.16
K084	phrase	사진 보다	1.13	사진 보다 (sajin boda)	ㄴ	ㅅ→ㅏ, ㅏ→ㅈ ㅈ→ㅣ, ㅣ→ㅂ ㅂ→ㅗ, ㅗ→ㄷ ㄷ→ㅏ		0.16	0.09	0	0.03	0.06
K085	phrase	노래 듣다	1.03	노래 듣따 (norae deuttta)	ㄷ	ㄴ→ㅗ, ㅗ→ㄹ ㄹ→ㅐ, ㅐ→ㄷ ㄷ→ㅡ ㅡ→ㄷ, ㄷ→ㄸ ㄸ→ㅏ	tensification	0.16	0.09	0.13	0.03	0.10
K086	phrase	영화 보다	1.03	영화 보다 (yeonghwa boda)	ㅇ	ㅇ→ㅕ, ㅕ→ㅇ ㅇ→ㅎ, ㅎ→ㅘ ㅘ→ㅂ ㅂ→ㅗ, ㅗ→ㄷ ㄷ→ㅏ		0.16	0.09	0	0.03	0.07

Continued on next page

K087	sentence	오늘 오후에는 친구와 근처 공원에서 산책을 했다.	3.87	오늘 오후에는 친구와 근처 공워네서 산채글 핻따 (oneul ohueneun chinguwa geuncheo gongwoneseo sanchaegeul haettta)	ㄹ,ㄴ,ㅇ,ㄷ	ㅇ→ㅗ, ㅗ→ㄴ ㄴ→ㅡ, ㅡ→ㄹ ㅇ→ㅗ, ㅗ→ㅎ ㅎ→ㅜ, ㅜ→ㅇ ㅇ→ㅖ, ㅖ→ㄴ ㄴ→ㅡ, ㅡ→ㄴ ㅊ→ㅣ, ㅣ→ㄴ ㄴ→ㄱ, ㄱ→ㅜ ㅜ→ㅇ, ㅇ→ㅘ ㄱ→ㅡ, ㅡ→ㄴ ㄴ→ㅊ, ㅊ→ㅓ ㄱ→ㅗ, ㅗ→ㅇ ㅇ→ㅇ, ㅇ→ㅝ ㅝ→ㄴ, ㄴ→ㅖ ㅖ→ㅅ, ㅅ→ㅓ ㅅ→ㅏ, ㅏ→ㄴ ㄴ→ㅊ, ㅊ→ㅐ ㅐ→ㄱ, ㄱ→ㅡ ㅡ→ㄹ, ㅎ→ㅐ ㅐ→ㄷ,ㄷ→ㅃ ㅃ→ㅏ	resyllabification, tensification	0.38	0.36	0.25	0.11	0.07
K088	sentence	아침에 일어나 창문을 열고 밖을 바라봤다.	3.07	아치메 이러나 창무늘 열고 바랄봗따 (achime ireona changmuneul yeolgo bakkeul barabwattta)	ㅇ,ㄹ,ㄷ	ㅇ→ㅏ, ㅏ→ㅊ ㅊ→ㅣ, ㅣ→ㅁ ㅁ→ㅖ ㅇ→ㅣ, ㅣ→ㄹ ㄹ→ㅓ, ㅓ→ㄴ ㄴ→ㅏ ㅊ→ㅏ, ㅏ→ㅇ ㅇ→ㅁ, ㅁ→ㅜ ㅜ→ㄴ, ㄴ→ㅡ ㅡ→ㄹ ㅇ→ㅕ, ㅕ→ㄹ ㄹ→ㄱ, ㄱ→ㅗ ㅂ→ㅏ, ㅏ→ㅍ ㅍ→ㅡ, ㅡ→ㄹ ㅂ→ㅏ, ㅏ→ㄹ ㄹ→ㅏ, ㅏ→ㅂ ㅂ→ㅘ, ㅘ→ㄷ ㄷ→ㅍ, ㅍ→ㅏ ㅎ→ㅏ, ㅏ→ㄱ ㄱ→ㅍ, ㅍ→ㅛ ㅛ→ㄱ, ㄱ→ㅏ ㅍ→ㅡ, ㅡ→ㄴ ㄴ→ㄴ, ㄴ→ㅏ	resyllabification, tensification	0.38	0.55	0.25	0.10	0.10
K089	sentence	학교가 끝나고 친구와 같이 집에 돌아왔다.	3.3	학꾜가 끈나고 친구와 가치 지베 도라왇따 (hakkkyoga kkeunnago chinguwa gachi jibe dorawattta)	ㄱ,ㄴ,ㄷ	ㅏ→ㄱ, ㄱ→ㅛ ㅊ→ㅣ, ㅣ→ㄴ ㄴ→ㄱ, ㄱ→ㅜ ㅜ→ㅇ, ㅇ→ㅘ ㄱ→ㅏ, ㅏ→ㅊ ㅊ→ㅣ ㅈ→ㅣ, ㅣ→ㅂ ㅂ→ㅖ ㄷ→ㅗ, ㅗ→ㄹ ㄹ→ㅏ, ㅏ→ㅇ ㅇ→ㅘ, ㅘ→ㄷ ㄷ→ㅍ, ㅍ→ㅏ ㅈ→ㅓ, ㅓ→ㅁ ㅁ→ㅅ, ㅅ→ㅣ ㅣ→ㅁ, ㅁ→ㅖ ㅖ→ㄴ, ㄴ→ㅡ ㅡ→ㄴ	tensification; nasal assimilation palatalization; resyllabification	0.38	0.36	0.5	0.09	0.10
K090	sentence	점심에는 따뜻한 국물과 밥을 함께 먹었다.	3.23	점시메는 따뜨탄 궁물과 바블 함께 머걷따 (jeomsimeneun ttatteutan gungmulgwa babeul hamkke meogeottta)	ㅁ,ㄴ,ㅇ,ㄹ,ㄷ	ㅍ→ㅏ, ㅏ→ㅍ ㅍ→ㅡ, ㅡ→ㅌ ㅌ→ㅏ, ㅏ→ㄴ ㄱ→ㅜ, ㅜ→ㅇ ㅇ→ㅁ, ㅁ→ㅜ ㅜ→ㄹ, ㄹ→ㄱ ㄱ→ㅘ ㅂ→ㅏ, ㅏ→ㅂ ㅂ→ㅡ, ㅡ→ㄹ ㅎ→ㅏ, ㅏ→ㅁ ㅁ→ㅠ, ㅠ→ㅖ ㅁ→ㅓ, ㅓ→ㄱ ㄱ→ㅓ, ㅓ→ㄷ ㄷ→ㅍ, ㅍ→ㅏ	resyllabification, aspiration; nasal assimilation tensification	0.4	0.73	0.5	0.11	0.13

Continued on next page

K091	sentence	오늘 저녁에는 뭐 먹을까?	2	오늘 저녀게는 뭐 머글까 (oneul jeonyeogeneun mwo meogeulkka)	ㄹ, ㄴ	ㅇ→ㅗ, ㅗ→ㄴ ㄴ→ㅡ,ㅡ→ㄹ, ㄹ→ㅈ, ㅈ→ㅓ, ㅕ→ㄴ,ㄴ→ㅖ, ㅖ→ㄱ, ㄱ→ㅖ ㄴ→ㅡ,ㅡ→ㄴ ㅁ→ㅝ, ㅁ→ㅓ, ㅓ→ㄱ,ㄱ→ㅡ, ㅡ→ㄹ, ㄹ→ㅂ ㅂ→ㅏ	resyllabification	0.32	0.36	0.13	0.05	0.13
K092	sentence	잠깐만, 물 좀 마시고 가자.	2.23	잠깐만 물 좀 마시고 가자 (jamkkanman mul jom masigo gaja)	ㅁ, ㄴ, ㄹ	ㅈ→ㅏ, ㅏ→ㅁ ㅁ→ㅂ, ㅂ→ㅏ ㅏ→ㄴ, ㄴ→ㅁ ㅁ→ ㅏ, ㅏ→ㄴ ㅁ→ㅜ, ㅜ→ㄹ ㅈ→ㅗ, ㅗ→ㅁ ㅁ→ ㅏ, ㅏ→ㅅ ㅅ→ ㅣ, ㅣ→ㄱ ㄱ→ㅗ ㄱ→ ㅏ, ㅏ→ㅈ ㅈ→ㅏ		0.22	0.27	0	0.05	0.06
K093	sentence	주말에는 친구를 만나 영화도 보고 밥도 먹었다.	3.47	주마레는 친구를 만나 영화도 보고 밥또 머걷따 (jumareneun chingureul manna yeonghwado bogo baptto meogeottta)	ㄴ, ㄹ, ㅇ, ㅂ, ㄷ	ㅈ→ㅜ, ㅜ→ㅁ ㅁ→ ㅏ, ㅏ→ㄹ ㄹ→ㅖ, ㅖ→ㄴ ㄴ→ㅡ, ㅡ→ㄴ ㅊ→ ㅣ, ㅣ→ㄴ ㄴ→ㄱ, ㄱ→ㅜ ㅜ→ㄹ, ㄹ→ㅡ ㅡ→ㄹ ㅁ→ ㅏ, ㅏ→ㄴ ㄴ→ㄴ, ㄴ→ ㅏ ㅇ→ ㅕ, ㅕ→ㅇ ㅇ→ㅎ, ㅎ→ㅘ ㅘ→ㄷ, ㄷ→ㅗ ㅂ→ㅗ, ㅗ→ㄱ ㄱ→ㅗ ㅂ→ ㅏ, ㅏ→ㅂ ㅂ→ㅃ, ㅃ→ㅗ ㅁ→ ㅓ, ㅓ→ㄱ ㄱ→ ㅓ, ㅓ→ㄷ ㄷ→ㅃ, ㅃ→ ㅏ ㅇ→ ㅣ	resyllabification, tensification	0.4	0.55	0.25	0.11	0.09
K094	sentence	이 옷보다 파란색이 더 잘 어울리는 것 같아.	2.83	이 옫뽀다 파란새기 더 잘 어울리는 걷 까타 (i otppoda paransaegi deo jal eoullineun geot kkata)	ㄷ, ㄴ, ㄹ	ㅇ→ㅗ, ㅗ→ㄷ ㄷ→ㅃ, ㅃ→ㅗ ㅗ→ㄷ, ㄷ→ ㅏ ㅍ→ ㅏ, ㅏ→ㄹ ㄹ→ ㅏ, ㅏ→ㄴ ㄴ→ㅅ, ㅅ→ ㅐ ㅐ→ㄱ, ㄱ→ ㅣ ㄷ→ ㅓ ㅈ→ ㅏ, ㅏ→ㄹ ㅇ→ ㅓ, ㅓ→ㅇ ㅇ→ㅜ, ㅜ→ㄹ ㄹ→ㄹ, ㄹ→ ㅣ ㅣ→ㄴ, ㄴ→ㅡ ㅡ→ㄴ ㄱ→ ㅓ, ㅓ→ㄷ ㄲ→ ㅏ, ㅏ→ㅌ ㅌ→ ㅏ ㅈ→ ㅏ, ㅏ→ㄱ ㄱ→ ㅣ	tensification; resyllabification, aspiration	0.36	0.36	0.38	0.09	0.10
K095	sentence	자기 전에 물을 마시고 책을 조금 읽었다.	2.88	자기 저네 무를 마시고 채글 쪼금 일걷따 (jagi jeone mureul masigo chaegeul jjogeum ilgeottta)	ㄹ, ㅁ, ㄷ	ㅈ→ㅓ, ㅓ→ㄴ ㄴ→ㅖ ㅁ→ㅜ, ㅜ→ㄹ ㄹ→ㅡ, ㅡ→ㄹ ㅁ→ ㅏ, ㅏ→ㅅ ㅅ→ ㅣ, ㅣ→ㄱ ㄱ→ㅗ ㅊ→ㅐ, ㅐ→ㄱ ㄱ→ㅡ, ㅡ→ㄹ ㅉ→ㅗ, ㅗ→ㄱ ㄱ→ㅡ, ㅡ→ㅁ ㅇ→ ㅣ, ㅣ→ㄹ ㄹ→ㄱ, ㄱ→ ㅓ ㅓ→ㄷ, ㄷ→ㅍ ㅍ→ ㅏ	resyllabification, tensification	0.38	0.36	0.25	0.08	0.09

Continued on next page

K096	sentence	아, 맞다. 내일 같이 학교 가기로 했지?	2.7	아 맏따 내일 가치 학꾜 가기로 핻찌 (a mattta naeil gachi hakkkyo gagiro haetjji)	ㄷ, ㄹ, ㄱ	ㅇ → ㅏ ㅁ → ㅏ, ㅏ → ㄷ ㄷ → ㄸ, ㄸ → ㅏ ㄴ → ㅐ, ㅐ → ㅇ ㅇ → ㅣ, ㅣ → ㄹ ㄱ → ㅏ, ㅏ → ㅊ ㅊ → ㅣ ㅎ → ㅏ, ㅏ → ㄱ ㄱ → ㄲ, ㄲ → ㅆ ㄱ → ㅏ, ㅏ → ㄱ ㄱ → ㅣ, ㅣ → ㄹ ㄹ → ㅗ ㅎ → ㅐ, ㅐ → ㄷ ㄷ → ㅉ, ㅉ → ㅣ	tensification; palatalization	0.32	0.27	0.25	0.06	0.08

Table 29: Full Japanese lexicon table ($N = 96$), consisting of words and phrases. The table reports mean actual duration (Dur.), canonical pronunciation (Pron.), phoneme transitions (Tr), phonological processes (Ph), prosodic features (Pros), phoneme coverage (C_{ph}), prosodic coverage (C_{pros}), phoneme-transition coverage (C_{ptr}), and audio utility (U_A).

ID	Type	Text	Kana	Dur.	Pron.	Tr	Ph	Pros	C_{ph}	C_{pros}	C_{ptr}	U_A
J001	word	朝	あさ	0.52	あさ (asa)	a→s, s→a		mora timed, pitch accent relevant	0.06	0.00	0.01	0.02
J002	word	家	いえ	0.52	いえ (ie)	i→e		mora timed, pitch accent relevant	0.06	0.00	0.01	0.02
J003	word	上	うえ	0.52	うえ (ue)	u→e		mora timed, pitch accent relevant	0.06	0.00	0.01	0.02
J004	word	駅	えき	0.52	えき (eki)	e→k, k→i		mora timed, pitch accent relevant	0.09	0.00	0.01	0.03
J005	word	丘	おか	0.52	おか (oka)	o→k, k→a		mora timed, pitch accent relevant	0.09	0.00	0.01	0.03
J006	word	柿	かき	0.52	かき (kaki)	k→a, a→k k→i		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J007	word	鍵	かぎ	0.52	かぎ (kagi)	k→a, a→g g→i		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J008	word	傘	かさ	0.52	かさ (kasa)	k→a, a→s s→a		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J009	word	風	かぜ	0.52	かぜ (kaze)	k→a, a→z z→e		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J010	word	肩	かた	0.52	かた (kata)	k→a, a→t t→a		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J011	word	角	かど	0.52	かど (kado)	k→a, a→d d→o		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J012	word	壁	かべ	0.52	かべ (kabe)	k→a, a→b b→e		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J013	word	紙	かみ	0.52	かみ (kami)	k→a, a→m m→i		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J014	word	殻	から	0.52	から (kara)	k→a, a→r r→a		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J015	word	川	かわ	0.52	かわ (kawa)	k→a, a→w w→a		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J016	word	菊	きく	0.52	きく (kiku)	k→i, i→k k→u	high vowel devoicing context	mora timed, pitch accent relevant	0.09	0.20	0.02	0.08
J017	word	靴	くつ	0.52	くつ (kutsu)	k→u, u→ts ts→u	high vowel devoicing context	mora timed, pitch accent relevant	0.09	0.20	0.02	0.08
J018	word	雲	くも	0.52	くも (kumo)	k→u, u→m m→o		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J019	word	毛	け	0.33	け (ke)	k→e		mora timed, pitch accent relevant	0.06	0.00	0.01	0.02
J020	word	声	こえ	0.52	こえ (koe)	k→o, o→e		mora timed, pitch accent relevant	0.09	0.00	0.01	0.03
J021	word	ここ	ここ	0.52	ここ (koko)	k→o, o→k k→o		mora timed, pitch accent relevant	0.06	0.00	0.01	0.02

Continued on next page

J022	word	坂	さか	0.52	さか (saka)	s→a, a→k k→a		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J023	word	寿司	すし	0.52	すし (sushi)	s→u, u→sh sh→i	high vowel devoicing context	mora timed, pitch accent relevant	0.12	0.20	0.02	0.09
J024	word	外	そと	0.52	そと (soto)	s→o, o→t t→o		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J025	word	滝	たき	0.52	たき (taki)	t→a, a→k k→i		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J026	word	地図	ちず	0.52	ちず (chizu)	ch→i, i→z z→u		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J027	word	月	つき	0.52	つき (tsuki)	ts→u, u→k k→i	high vowel devoicing context	mora timed, pitch accent relevant	0.12	0.20	0.02	0.09
J028	word	手	て	0.33	て (te)	t→e		mora timed, pitch accent relevant	0.06	0.00	0.01	0.02
J029	word	音	おと	0.52	おと (oto)	o→t, t→o		mora timed, pitch accent relevant	0.06	0.00	0.01	0.02
J030	word	夏	なつ	0.52	なつ (natsu)	n→a, a→ts ts→u		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J031	word	庭	にわ	0.52	にわ (niwa)	n→i, i→w w→a		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J032	word	布	ぬの	0.52	ぬの (nuno)	n→u, u→n n→o		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J033	word	猫	ねこ	0.52	ねこ (neko)	n→e, e→k k→o		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J034	word	花	はな	0.52	はな (hana)	h→a, a→n n→a		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J035	word	人	ひと	0.52	ひと (hito)	h→i, i→t t→o	high vowel devoicing context	mora timed, pitch accent relevant	0.12	0.20	0.02	0.09
J036	word	船	ふね	0.52	ふね (fune)	f→u, u→n n→e		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J037	word	部屋	へや	0.52	へや (heya)	h→e, e→y y→a		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J038	word	星	ほし	0.52	ほし (hoshi)	h→o, o→sh sh→i		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J039	word	窓	まど	0.52	まど (mado)	m→a, a→d d→o		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J040	word	水	みず	0.52	みず (mizu)	m→i, i→z z→u		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J041	word	胸	むね	0.52	むね (mune)	m→u, u→n n→e		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J042	word	目	め	0.33	め (me)	m→e		mora timed, pitch accent relevant	0.06	0.00	0.01	0.02
J043	word	森	もり	0.52	もり (mori)	m→o, o→r r→i		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J044	word	山	やま	0.52	やま (yama)	y→a, a→m m→a		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J045	word	雪	ゆき	0.52	ゆき (yuki)	y→u, u→k k→i		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04

Continued on next page

J046	word	夜	よる	0.52	よる (yoru)	y→o, o→r r→u		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J047	word	空	そら	0.52	そら (sora)	s→o, o→r r→a		mora timed, pitch accent relevant	0.12	0.00	0.02	0.04
J048	word	林檎	りんご	0.74	りんご (ringo)	r→i, i→N N→g, g→o	moraic_nasal	mora timed, pitch accent relevant	0.15	0.20	0.03	0.18
J049	word	留守	るす	0.52	るす (rusu)	r→u, u→s s→u		mora timed, pitch accent relevant	0.09	0.00	0.02	0.03
J050	word	冷蔵庫	れいぞうこ	1.14	れいぞうこ (reizouko)	r→e, e→i i→z, z→o o→u, u→k k→o	vowel length	mora timed, pitch accent relevant	0.21	0.20	0.05	0.20
J051	word	廊下	ろうか	0.74	ろうか (rouka)	r→o, o→u u→k, k→a	vowel length	mora timed, pitch accent relevant	0.15	0.20	0.03	0.18
J052	word	私	わたし	0.71	わたし (watashi)	w→a, a→t t→a, a→sh sh→i		mora timed, pitch accent relevant	0.15	0.00	0.03	0.05
J053	word	銀行	ぎんこう	0.96	ぎんこう (ginkou)	g→i, i→N N→k, k→o o→u	moraic_nasal vowel length	mora timed, pitch accent relevant	0.18	0.40	0.03	0.32
J054	word	午後	ごご	0.52	ごご (gogo)	g→o, o→g g→o		mora timed, pitch accent relevant	0.06	0.00	0.01	0.02
J055	word	写真	しゃしん	0.74	しゃしん (shashin)	sh→a, a→sh sh→i, i→N	moraic_nasal palatalization	mora timed, pitch accent relevant	0.12	0.40	0.03	0.22
J056	word	主役	しゅやく	0.71	しゅやく (shuyaku)	sh→u, u→y y→a, a→k k→u	palatalization	mora timed, pitch accent relevant	0.15	0.20	0.03	0.10
J057	word	食事	しょくじ	0.71	しょくじ (shokuji)	sh→o, o→k k→u, u→j j→i	palatalization	mora timed, pitch accent relevant	0.18	0.20	0.03	0.10
J058	word	茶碗	ちゃわん	0.74	ちゃわん (chawan)	ch→a, a→w w→a, a→N	moraic_nasal palatalization	mora timed, pitch accent relevant	0.12	0.40	0.03	0.22
J059	word	注意	ちゅうい	0.74	ちゅうい (chuui)	ch→u, u→u u→i	vowel length palatalization	mora timed, pitch accent relevant	0.09	0.40	0.02	0.21
J060	word	貯金	ちょきん	0.74	ちょきん (chokin)	ch→o, o→k k→i, i→N	moraic_nasal palatalization	mora timed, pitch accent relevant	0.15	0.40	0.03	0.23
J061	word	客	きゃく	0.52	きゃく (kyaku)	ky→a, a→k k→u	palatalization	mora timed, pitch accent relevant	0.12	0.20	0.02	0.09
J062	word	九	きゅう	0.55	きゅう (kyuu)	ky→u, u→u	vowel length palatalization	mora timed, pitch accent relevant	0.06	0.40	0.01	0.20
J063	word	今日	きょう	0.55	きょう (kyou)	ky→o, o→u	vowel length palatalization	mora timed, pitch accent relevant	0.09	0.40	0.01	0.21
J064	word	百	ひゃく	0.52	ひゃく (hyaku)	hy→a, a→k k→u	palatalization	mora timed, pitch accent relevant	0.12	0.20	0.02	0.09
J065	word	表	ひょう	0.55	ひょう (hyou)	hy→o, o→u	vowel length palatalization	mora timed, pitch accent relevant	0.09	0.40	0.01	0.21
J066	word	牛乳	ぎゅうにゅう	0.95	ぎゅうにゅう (gyuunyuu)	gy→u, u→u u→ny, ny→u u→u	vowel length palatalization	mora timed, pitch accent relevant	0.09	0.40	0.03	0.21
J067	word	料理	りょうり	0.74	りょうり (ryouri)	ry→o, o→u u→r, r→i	vowel length palatalization	mora timed, pitch accent relevant	0.15	0.40	0.03	0.23
J068	word	旅行	りょこう	0.74	りょこう (ryokou)	ry→o, o→k k→o, o→u	vowel length palatalization	mora timed, pitch accent relevant	0.12	0.40	0.03	0.22
J069	word	病院	びょういん	0.96	びょういん (byouin)	by→o, o→u u→i, i→N	moraic_nasal vowel length palatalization	mora timed, pitch accent relevant	0.15	0.60	0.03	0.36

Continued on next page

J070	word	発表	はっぴょう	0.97	はっぴょう (happyou)	h→a, a→Q Q→py, py→o o→u	gemination vowel length palatalization	mora timed, pitch accent relevant	0.18	0.60	0.03	0.37
J071	word	学校	がっこう	0.97	がっこう (gakkou)	g→a, a→Q Q→k, k→o o→u	gemination vowel length	mora timed, pitch accent relevant	0.18	0.40	0.03	0.32
J072	word	切手	きって	0.76	きって (kitte)	k→i, i→Q Q→t, t→e	gemination	mora timed, pitch accent relevant	0.15	0.20	0.03	0.18
J073	word	雑誌	ざっし	0.76	ざっし (zasshi)	z→a, a→Q Q→sh, sh→i	gemination	mora timed, pitch accent relevant	0.15	0.20	0.03	0.18
J074	word	一杯	いっぱい	0.95	いっぱい (ippai)	i→Q, Q→p p→a, a→i	gemination	mora timed, pitch accent relevant	0.12	0.20	0.03	0.17
J075	word	新聞	しんぶん	0.93	しんぶん (shinbun)	sh→i, i→N N→b, b→u u→N	moraic_nasal	mora timed, pitch accent relevant	0.15	0.20	0.03	0.18
J076	word	安全	あんぜん	0.93	あんぜん (anzen)	a→N, N→z z→e, e→N	moraic_nasal	mora timed, pitch accent relevant	0.12	0.20	0.03	0.17
J077	word	案内	あんない	0.93	あんない (annai)	a→N, N→n n→a, a→i	moraic_nasal	mora timed, pitch accent relevant	0.12	0.20	0.03	0.17
J078	word	温泉	おんせん	0.93	おんせん (onsen)	o→N, N→s s→e, e→N	moraic_nasal	mora timed, pitch accent relevant	0.12	0.20	0.03	0.17
J079	word	先生	せんせい	0.96	せんせい (sensei)	s→e, e→N N→s, s→e e→i	moraic_nasal vowel length	mora timed, pitch accent relevant	0.12	0.40	0.03	0.30
J080	word	東京	とうきょう	0.95	とうきょう (toukyou)	t→o, o→u u→ky, ky→o o→u	vowel length palatalization	mora timed, pitch accent relevant	0.12	0.40	0.03	0.22
J081	phrase	青い空	あおいそら	1.17	あおいそら (aoisora)	a→o, o→i i→s, s→o o→r, r→a		mora timed, pitch accent relevant	0.15	0.00	0.04	0.05
J082	phrase	赤い傘	あかいかさ	1.17	あかいかさ (akaikasa)	a→k, k→a a→i, i→k k→a, a→s s→a		mora timed, pitch accent relevant	0.12	0.00	0.04	0.04
J083	phrase	大きな木	おおきなき	1.19	おおきなき (ookinaki)	o→o, o→k k→i, i→n n→a, a→k k→i	vowel length	mora timed, pitch accent relevant	0.15	0.20	0.04	0.18
J084	phrase	小さな猫	ちいさなねこ	1.39	ちいさなねこ (chiisananeko)	ch→i, i→i i→s, s→a a→n, n→a a→n, n→e e→k, k→o	vowel length	mora timed, pitch accent relevant	0.24	0.20	0.06	0.21
J085	phrase	白い雲	しろいくも	1.17	しろいくも (shiroikumo)	sh→i, i→r r→o, o→i i→k, k→u u→m, m→o		mora timed, pitch accent relevant	0.21	0.00	0.05	0.07
J086	phrase	黒い靴	くろいくつ	1.17	くろいくつ (kuroikutsu)	k→u, u→r r→o, o→i i→k, k→u u→ts, ts→u	high vowel devoicing context	mora timed, pitch accent relevant	0.18	0.20	0.05	0.11
J087	phrase	新しい地図	あたらしいちず	1.58	あたらしいちず (atarashiichizu)	a→t, t→a a→r, r→a a→sh, sh→i i→i, i→ch ch→i, i→z z→u	vowel length	mora timed, pitch accent relevant	0.24	0.20	0.07	0.21
J088	phrase	静かな部屋	しずかなへや	1.36	しずかなへや (shizukanaheya)	sh→i, i→z z→u, u→k k→a, a→n n→a, a→h h→e, e→y y→a		mora timed, pitch accent relevant	0.30	0.00	0.07	0.09
J089	phrase	きれいな花	きれいなはな	1.39	きれいなはな (kireinahana)	k→i, i→r r→e, e→i i→n, n→a a→h, h→a a→n, n→a	vowel length	mora timed, pitch accent relevant	0.21	0.20	0.06	0.20

Continued on next page

J090	phrase	冷たい水	つめたい みず	1.36	つめたいみず (tsumetaimizu)	ts→u, u→m m→e, e→t t→a, a→i i→m, m→i i→z, z→u		mora timed, pitch accent relevant	0.24	0.00	0.07	0.08
J091	phrase	温かいお茶	あたたかい おちゃ	1.55	あたたかいおちゃ (atatakai ocha)	a→t, t→a a→t, t→a a→k, k→a a→i, i→o o→ch, ch→a	palatalization	mora timed, pitch accent relevant	0.18	0.20	0.05	0.11
J092	phrase	切符を買う	きっぷを かう	1.41	きっぷをかう (kippu o kau)	k→i, i→Q Q→p, p→u u→o, o→k k→a, a→u	gemination	mora timed, pitch accent relevant	0.21	0.20	0.05	0.20
J093	phrase	写真を撮る	しゃしんを とる	1.39	しゃしんをとる (shashin o toru)	sh→a, a→sh sh→i, i→N N→o, o→t t→o, o→r r→u	moraic_nasal palatalization	mora timed, pitch accent relevant	0.24	0.40	0.06	0.26
J094	phrase	ゆっくり話す	ゆっくり はなす	1.60	ゆっくりはなす (yukkuri hanasu)	y→u, u→Q Q→k, k→u u→r, r→i i→h, h→a a→n, n→a a→s, s→u	gemination	mora timed, pitch accent relevant	0.30	0.20	0.08	0.23
J095	phrase	駅まで歩く	えきまで あるく	1.55	えきまであるく (eki made aruku)	e→k, k→i i→m, m→a a→d, d→e e→a, a→r r→u, u→k k→u		mora timed, pitch accent relevant	0.24	0.00	0.07	0.08
J096	phrase	今日は晴れです	きょうわ はれです	1.58	きょうわはれです (kyou wa hare desu)	ky→o, o→u u→w, w→a a→h, h→a a→r, r→e e→d, d→e e→s, s→u	vowel length palatalization	mora timed, pitch accent relevant	0.30	0.40	0.08	0.28

Table 30: Full Vietnamese lexicon table ($N = 96$), consisting of words and phrases. The table reports mean actual duration (Dur.), pronunciation (Pron.), coda inventory (Co), phoneme transitions (Tr), tone categories (Tone), phonological features (Ph), coda coverage gain (C_{coda}), tone coverage gain (C_{tone}), phoneme-transition coverage gain (C_{ptr}), and audio utility (U_A).

ID	Type	Text	Dur.	Pron.	Co	Tr	Tone	Ph	C_{coda}	C_{tone}	C_{ptr}	U_A
V001	word	ba	0.52	ɓaː (bah)	∅	ɓ→aː	ngang		0.00	0.17	0.01	0.05
V002	word	bà	0.52	ɓaː (bah)	∅	ɓ→aː	huyền		0.00	0.17	0.01	0.05
V003	word	cá	0.52	kaː (kah)	∅	k→aː	sắc		0.00	0.17	0.01	0.05
V004	word	cả	0.52	kaː (kah)	∅	k→aː	hỏi		0.00	0.17	0.01	0.05
V005	word	mã	0.52	maː (mah)	∅	m→aː	ngã	phonation contrast	0.00	0.17	0.01	0.05
V006	word	bạn	0.52	ɓaːn (bahn)	n	ɓ→aː, aː→n	nặng	phonation contrast	0.08	0.17	0.01	0.08
V007	word	nhà	0.52	ɲaː (nyah)	∅	ɲ→aː	huyền		0.00	0.17	0.01	0.05
V008	word	hoa	0.52	hwaː (hwah)	∅	h→waː	ngang		0.00	0.17	0.01	0.05
V009	word	xe	0.52	sɛ (seh)	∅	s→ɛ	ngang		0.00	0.17	0.01	0.05
V010	word	đi	0.52	ɗi (dee)	∅	ɗ→i	ngang		0.00	0.17	0.01	0.05
V011	word	về	0.52	ve (veh)	∅	v→e	huyền		0.00	0.17	0.01	0.05
V012	word	ăn	0.52	an (ahn)	n	a→n	ngang		0.08	0.17	0.01	0.07
V013	word	cơm	0.52	kɤm (guhm)	m	k→ɤ, ɤ→m	ngang		0.08	0.17	0.01	0.08
V014	word	nước	0.49	nɯək (noo-uhk)	c	n→ɯə, ɯə→k	sắc	checked syllable	0.08	0.17	0.01	0.08
V015	word	trà	0.52	ʈaː (trah)	∅	ʈ→aː	huyền		0.00	0.17	0.01	0.05
V016	word	sữa	0.52	ʂɯə (shoo-uh)	∅	ʂ→ɯə	ngã	phonation contrast	0.00	0.17	0.01	0.05
V017	word	sách	0.49	ʂaːc (shahk)	ch	ʂ→aː, aː→c	sắc	checked syllable	0.08	0.17	0.01	0.08
V018	word	bút	0.49	ɓut (boot)	t	ɓ→u, u→t	sắc	checked syllable	0.08	0.17	0.01	0.08
V019	word	bàn	0.52	ɓaːn (bahn)	n	ɓ→aː, aː→n	huyền		0.08	0.17	0.01	0.08
V020	word	ghế	0.52	ɣe (gheh)	∅	ɣ→e	sắc		0.00	0.17	0.01	0.05
V021	word	cửa	0.52	kɯə (koo-uh)	∅	k→ɯə	hỏi		0.00	0.17	0.01	0.05
V022	word	biển	0.52	ɓiən (bee-uhn)	n	ɓ→iə, iə→n	hỏi		0.08	0.17	0.01	0.08
V023	word	trời	0.52	ʈɤj (truhy)	i	ʈ→ɤ, ɤ→j	huyền		0.08	0.17	0.01	0.08
V024	word	đất	0.49	ɗət (duht)	t	ɗ→ə, ə→t	sắc	checked syllable	0.08	0.17	0.01	0.08
V025	word	núi	0.52	nuj (nooy)	i	n→u, u→j	sắc		0.08	0.17	0.01	0.08
V026	word	sông	0.52	ʂoŋ (shohng)	ng	ʂ→o, o→ŋ	ngang		0.08	0.17	0.01	0.08
V027	word	cây	0.52	kəj (kuhy)	y	k→ə, ə→j	ngang		0.08	0.17	0.01	0.08
V028	word	quả	0.52	kwaː (kwah)	∅	kw→aː	hỏi		0.00	0.17	0.01	0.05
V029	word	áo	0.52	aːw (ow)	o	aː→w	sắc		0.08	0.17	0.01	0.07
V030	word	quần	0.52	kwən (kwuhn)	n	kw→ə, ə→n	huyền		0.08	0.17	0.01	0.08
V031	word	mắt	0.49	mat (maht)	t	m→a, a→t	sắc	checked syllable	0.08	0.17	0.01	0.08
V032	word	mũi	0.52	muj (mooy)	i	m→u, u→j	ngã	phonation contrast	0.08	0.17	0.01	0.08
V033	word	tóc	0.49	tɔk (tawk)	c	t→ɔ, ɔ→k	sắc	checked syllable	0.08	0.17	0.01	0.08
V034	word	tay	0.52	taːj (tahy)	y	t→aː, aː→j	ngang		0.08	0.17	0.01	0.08
V035	word	chân	0.52	cən (chuhn)	n	c→ə, ə→n	ngang		0.08	0.17	0.01	0.08

Continued on next page

ID	Type	Text	Dur.	Pron.	Co	Tr	Tone	Ph	C_{coda}	C_{tone}	C_{ptr}	U_A
V036	word	mặt	0.49	mat (maht)	t	m→a, a→t	nặng	checked syllable, phonation contrast	0.08	0.17	0.01	0.08
V037	word	miệng	0.52	miəŋ (mee-uhng)	ng	m→iə, iə→ŋ	nặng	phonation contrast	0.08	0.17	0.01	0.08
V038	word	tai	0.52	taːj (tahy)	i	t→aː, aː→j	ngang		0.08	0.17	0.01	0.08
V039	word	mẹ	0.52	mɛ (meh)	∅	m→ɛ	nặng	phonation contrast	0.00	0.17	0.01	0.05
V040	word	bố	0.52	ɓo (boh)	∅	ɓ→o	sắc		0.00	0.17	0.01	0.05
V041	word	anh	0.52	aːɲ (ahny)	nh	aː→ɲ	ngang		0.08	0.17	0.01	0.07
V042	word	chị	0.52	ci (chee)	∅	c→i	nặng	phonation contrast	0.00	0.17	0.01	0.05
V043	word	em	0.52	ɛm (em)	m	ɛ→m	ngang		0.08	0.17	0.01	0.07
V044	word	người	0.52	ŋɯəj (ngoo-uhy)	i	ŋ→ɯə, ɯə→j	huyền		0.08	0.17	0.01	0.08
V045	word	trẻ	0.52	ɻɛ (treh)	∅	ɻ→ɛ	hỏi		0.00	0.17	0.01	0.05
V046	word	nhỏ	0.52	ɲɔ (nyaw)	∅	ɲ→ɔ	hỏi		0.00	0.17	0.01	0.05
V047	word	lớn	0.52	lʏn (luhn)	n	l→ʏ, ʏ→n	sắc		0.08	0.17	0.01	0.08
V048	word	đẹp	0.49	ɗɛp (dep)	p	ɗ→ɛ, ɛ→p	nặng	checked syllable, phonation contrast	0.08	0.17	0.01	0.08
V049	word	khỏe	0.52	xwɛ (khweh)	∅	x→wɛ	hỏi		0.00	0.17	0.01	0.05
V050	word	lạnh	0.52	laːɲ (lahny)	nh	l→aː, aː→ɲ	nặng	phonation contrast	0.08	0.17	0.01	0.08
V051	word	nóng	0.52	nɔŋ (nawng)	ng	n→ɔ, ɔ→ŋ	sắc		0.08	0.17	0.01	0.08
V052	word	sáng	0.52	ʂaːŋ (shahng)	ng	ʂ→aː, aː→ŋ	sắc		0.08	0.17	0.01	0.08
V053	word	tối	0.52	toj (toy)	i	t→o, o→j	sắc		0.08	0.17	0.01	0.08
V054	word	đỏ	0.52	ɗɔ (daw)	∅	ɗ→ɔ	hỏi		0.00	0.17	0.01	0.05
V055	word	trắng	0.52	ɻaŋ (trahng)	ng	ɻ→a, a→ŋ	sắc		0.08	0.17	0.01	0.08
V056	word	xanh	0.52	saːɲ (sahny)	nh	s→aː, aː→ɲ	ngang		0.08	0.17	0.01	0.08
V057	word	vàng	0.52	vaːŋ (vahng)	ng	v→aː, aː→ŋ	huyền		0.08	0.17	0.01	0.08
V058	word	đen	0.52	ɗɛn (den)	n	ɗ→ɛ, ɛ→n	ngang		0.08	0.17	0.01	0.08
V059	word	hỏi	0.52	hɔj (hoy)	i	h→ɔ, ɔ→j	hỏi		0.08	0.17	0.01	0.08
V060	word	nghĩ	0.52	ŋi (ngee)	∅	ŋ→i	ngã	phonation contrast	0.00	0.17	0.01	0.05
V061	word	giữ	0.52	zɯ (zoo)	∅	z→ɯ	ngã	phonation contrast	0.00	0.17	0.01	0.05
V062	word	đọc	0.49	ɗɔk (dawk)	c	ɗ→ɔ, ɔ→k	nặng	checked syllable, phonation contrast	0.08	0.17	0.01	0.08
V063	word	viết	0.49	viət (vee-uht)	t	v→iə, iə→t	sắc	checked syllable	0.08	0.17	0.01	0.08
V064	word	nói	0.52	nɔj (noy)	i	n→ɔ, ɔ→j	sắc		0.08	0.17	0.01	0.08
V065	word	nghe	0.52	ŋɛ (ngeh)	∅	ŋ→ɛ	ngang		0.00	0.17	0.01	0.05
V066	word	nhìn	0.52	ɲin (nyeen)	n	ɲ→i, i→n	huyền		0.08	0.17	0.01	0.08
V067	word	học	0.49	hɔk (hawk)	c	h→ɔ, ɔ→k	nặng	checked syllable, phonation contrast	0.08	0.17	0.01	0.08
V068	word	làm	0.52	laːm (lahm)	m	l→aː, aː→m	huyền		0.08	0.17	0.01	0.08
V069	word	chơi	0.52	cʏj (chuhy)	i	c→ʏ, ʏ→j	ngang		0.08	0.17	0.01	0.08
V070	word	ngủ	0.52	ŋu (ngoo)	∅	ŋ→u	hỏi		0.00	0.17	0.01	0.05
V071	word	trường	0.52	ɻɯəŋ (troo-uhng)	ng	ɻ→ɯə, ɯə→ŋ	huyền		0.08	0.17	0.01	0.08
V072	word	lớp	0.49	lʏp (luhp)	p	l→ʏ, ʏ→p	sắc	checked syllable	0.08	0.17	0.01	0.08
V073	word	việc	0.49	viək (vee-uhk)	c	v→iə, iə→k	nặng	checked syllable, phonation contrast	0.08	0.17	0.01	0.08
V074	word	bệnh	0.52	ɓeɲ (beny)	nh	ɓ→e, e→ɲ	nặng	phonation contrast	0.08	0.17	0.01	0.08

Continued on next page

ID	Type	Text	Dur.	Pron.	Co	Tr	Tone	Ph	C_{coda}	C_{tone}	C_{ptr}	U_A
V075	word	tiền	0.52	tiən (tee-uhn)	n	t→iə, iə→n	huyền		0.08	0.17	0.01	0.08
V076	word	đường	0.52	ɗɯəŋ (doo-uhng)	ng	ɗ→ɯə, ɯə→ŋ	huyền		0.08	0.17	0.01	0.08
V077	phrase	thành phố	0.94	tʰaːɲ fo (tahny foh)	nh,	tʰ→aː, aː→ɲ ɲ→f, f→o	huyền, sắc	tonal coarticulation context	0.08	0.33	0.03	0.14
V078	phrase	gia đình	0.94	zaː ɗiɲ (zah deeny)	nh,	z→aː, aː→ɗ ɗ→i, i→ɲ	huyền, ngang	tonal coarticulation context	0.08	0.33	0.03	0.14
V079	phrase	điện thoại	0.94	ɗiən tʰwaːj (dee-uhn twahy)	i, n	ɗ→iə, iə→n n→tʰ, tʰ→waː waː→j	nặng	phonation contrast, tonal coarticulation context	0.17	0.17	0.03	0.13
V080	phrase	máy tính	0.94	maːj tiɲ (mahy teeny)	nh, y	m→aː, aː→j j→t, t→i i→ɲ	sắc	tonal coarticulation context	0.17	0.17	0.03	0.13
V081	phrase	xin chào	0.94	sin caːw (seen chow)	n, o	s→i, i→n n→c, c→aː aː→w	huyền, ngang	tonal coarticulation context	0.17	0.33	0.03	0.17
V082	phrase	cảm ơn	0.94	kaːm ɤn (kahm uhn)	m, n	k→aː, aː→m m→ɤ, ɤ→n	hỏi, ngang	tonal coarticulation context	0.17	0.33	0.03	0.17
V083	phrase	chào buổi sáng	1.28	caːw ɓuəj ʂaːŋ (chow boo-uhy shahng)	i, ng o	c→aː, aː→w w→ɓ, ɓ→uə uə→j, j→ʂ ʂ→aː, aː→ŋ	huyền, hỏi sắc	tonal coarticulation context	0.25	0.50	0.05	0.26
V084	phrase	tôi tên là Lan	1.62	toj ten laː laːn (toy ten lah lahn)	i, n	t→o, o→j j→t, t→e e→n, n→l l→aː, aː→l l→aː, aː→n	huyền, ngang	tonal coarticulation context	0.17	0.33	0.06	0.20
V085	phrase	bạn khỏe không	1.28	ɓaːn xwɛ xoŋ (bahn khweh khohng)	n, ng	ɓ→aː, aː→n n→x, x→wɛ wɛ→x, x→o o→ŋ	hỏi, ngang nặng	phonation contrast, tonal coarticulatio context	0.17	0.50	0.04	0.23
V086	phrase	tôi uống nước	1.25	toj uəŋ nɯək (toy oo-uhng noo-uhk)	c, i ng	t→o, o→j j→uə, uə→ŋ ŋ→n, n→ɯə ɯə→k	ngang, sắc	checked syllable, tonal coarticulation context	0.25	0.33	0.04	0.21
V087	phrase	em đọc sách	1.22	ɛm ɗɔk ʂaːc (em dawk shahk)	c, ch m	ɛ→m, m→ɗ ɗ→ɔ, ɔ→k k→ʂ, ʂ→aː aː→c	ngang,nặng sắc	checked syllable, phonation contrast, tonal coarticulation context	0.25	0.50	0.04	0.26
V088	phrase	mẹ nấu cơm	1.28	mɛ nəw kɤm (meh nuh-oo guhm)	m, u	m→ɛ, ɛ→n n→ə, ə→w w→k, k→ɤ ɤ→m	ngang, nặng sắc	phonation contrast, tonal coarticulation context	0.17	0.50	0.04	0.23
V089	phrase	bố đi làm	1.28	ɓo ɗi laːm (boh dee lahm)	m,	ɓ→o, o→ɗ ɗ→i, i→l l→aː, aː→m	huyền, ngang sắc	tonal coarticulation context	0.08	0.50	0.04	0.21
V090	phrase	trẻ em chơi	1.28	ɻɛ ɛm cɤj (treh em chuhy)	i, m	ɻ→ɛ, ɛ→ɛ ɛ→m, m→c c→ɤ, ɤ→j	hỏi, ngang	tonal coarticulation context	0.17	0.33	0.04	0.18
V091	phrase	nhà ở đây	1.28	ɲaː ɤ ɗəj (nyah uh duhy)	y,	ɲ→aː, aː→ɤ ɤ→ɗ, ɗ→ə ə→j	huyền, hỏi ngang	tonal coarticulation context	0.08	0.50	0.03	0.20
V092	phrase	trời hôm nay đẹp	1.59	ʈɤj hom naːj ɗɛp (truhy hohm nahy dep)	i, m p, y	ʈ→ɤ, ɤ→j j→h, h→o o→m, m→n n→aː, aː→j j→ɗ, ɗ→ɛ ɛ→p	huyền, ngang nặng	checked syllable, phonation contrast, tonal coarticulation context	0.33	0.50	0.07	0.30
V093	phrase	tôi thích âm nhạc	1.56	toj tʰic ɔm ɲaːk (toy teek uhm nyahk)	c, ch i, m	t→o, o→j j→tʰ, tʰ→i i→c, c→ə ə→m, m→ɲ ɲ→aː, aː→k	ngang, nặng sắc	checked syllable, phonation contrast, tonal coarticulation context	0.33	0.50	0.06	0.30
V094	phrase	xin nói chậm	1.28	sin nɔj cəm (seen noy chuhm)	i, m n	s→i, i→n n→n, n→ɔ ɔ→j, j→c c→ə, ə→m	ngang, nặng sắc	phonation contrast, tonal coarticulation context	0.25	0.50	0.05	0.26
V095	phrase	hẹn gặp lại	1.25	hɛn ɣap laːj (hen gah lahy)	i, n p	h→ɛ, ɛ→n n→ɣ, ɣ→a a→p, p→l l→aː, aː→j	nặng	checked syllable, phonation contrast, tonal coarticulation context	0.25	0.17	0.05	0.17

Continued on next page

ID	Type	Text	Dur.	Pron.	Co	Tr	Tone	Ph	C_{coda}	C_{tone}	C_{ptr}	U_A
V096	phrase	Việt Nam rất đẹp	1.53	viət naːm zət dɛp (vee-uht nahm zuht dep)	m, p t	v→iə, iə→t t→n, n→aː aː→m, m→z z→ə, ə→t t→ɗ, ɗ→ɛ ɛ→p	ngang, nặng sắc	checked syllable, phonation contrast, tonal coarticulation context	0.25	0.50	0.07	0.28

Table 31: Full English lexicon table ($N = 96$), consisting of words and phrases. The table reports mean actual duration (Dur.), surface pronunciation (Pron.), consonant inventory (Cons.), vowel nuclei (Nuc.), phoneme transitions (Tr), phonological processes (Ph), phoneme coverage (C_{ph}), coda coverage (C_{coda}), stress-feature coverage (C_{stress}), and audio utility (U_A).

ID	Type	Text	Dur.	Pron.	Cons.	Nuc.	Tr	Ph	C_{ph}	C_{coda}	C_{stress}	U_A
E001	word	cat	0.48	kæt	k, t	æ	k→æ, æ→t	aspiration	0.07	0.06	0.33	0.08
E002	word	dog	0.48	dɔg	d, g	ɔ	d→ɔ, ɔ→g		0.07	0.06	0.33	0.07
E003	word	sheep	0.48	ʃiːp	ʃ, p	iː	ʃ→iː, iː→p		0.07	0.06	0.33	0.07
E004	word	fish	0.48	fɪʃ	f, ʃ	ɪ	f→ɪ, ɪ→ʃ		0.07	0.06	0.33	0.07
E005	word	bird	0.48	bɝd	b, d	ɝ	b→ɝ, ɝ→d	r coloring	0.07	0.06	0.33	0.10
E006	word	mouse	0.48	maʊs	m, s	aʊ	m→aʊ, aʊ→s		0.07	0.06	0.33	0.07
E007	word	goat	0.48	goʊt	g, t	oʊ	g→oʊ, oʊ→t		0.07	0.06	0.33	0.07
E008	word	zebra	0.80	ziːbɹə	z, b, ɹ	iː, ə	z→iː, iː→b b→ɹ, ɹ→ə	vowel reduction	0.12	0.00	0.67	0.15
E009	word	chair	0.48	tʃɛɹ	tʃ	ɛɹ	tʃ→ɛɹ	r coloring	0.05	0.00	0.33	0.09
E010	word	judge	0.48	dʒʊdʒ	dʒ	ʊ	dʒ→ʊ, ʊ→dʒ		0.05	0.06	0.33	0.06
E011	word	van	0.48	væn	v, n	æ	v→æ, æ→n		0.07	0.06	0.33	0.07
E012	word	think	0.52	θɪŋk	θ, ŋ, k	ɪ	θ→ɪ, ɪ→ŋ ŋ→k		0.09	0.06	0.33	0.08
E013	word	this	0.48	ðɪs	ð, s	ɪ	ð→ɪ, ɪ→s		0.07	0.06	0.33	0.07
E014	word	red	0.48	ɹɛd	ɹ, d	ɛ	ɹ→ɛ, ɛ→d		0.07	0.06	0.33	0.07
E015	word	light	0.48	laɪt	l, t	aɪ	l→aɪ, aɪ→t	dark l	0.07	0.06	0.33	0.08
E016	word	yellow	0.76	jɛloʊ	j, l	ɛ, oʊ	j→ɛ, ɛ→l l→oʊ	dark l	0.09	0.00	0.67	0.13
E017	word	water	0.76	wɔɹɝ	w, t	ɔ, ɝ	w→ɔ, ɔ→t t→ɝ	flapping, r coloring	0.09	0.00	0.67	0.16
E018	word	happy	0.76	hæpi	h, p	æ, i	h→æ, æ→p p→i		0.09	0.00	0.67	0.11
E019	word	pen	0.48	pɛn	p, n	ɛ	p→ɛ, ɛ→n	aspiration	0.07	0.06	0.33	0.08
E020	word	book	0.48	bʊk	b, k	ʊ	b→ʊ, ʊ→k		0.07	0.06	0.33	0.07
E021	word	cup	0.48	kʊp	k, p	ʊ	k→ʊ, ʊ→p	aspiration	0.07	0.06	0.33	0.08
E022	word	key	0.48	kiː	k	iː	k→iː	aspiration	0.05	0.00	0.33	0.07
E023	word	phone	0.48	foʊn	f, n	oʊ	f→oʊ, oʊ→n		0.07	0.06	0.33	0.07
E024	word	mirror	0.76	mɪɹɝ	m, ɹ	ɪ, ɝ	m→ɪ, ɪ→ɹ ɹ→ɝ	r coloring	0.09	0.00	0.67	0.14
E025	word	spoon	0.52	spuːn	s, p, n	uː	s→p, p→uː uː→n		0.09	0.06	0.33	0.08
E026	word	plate	0.52	pleɪt	p, l, t	eɪ	p→l, l→eɪ eɪ→t	aspiration	0.09	0.06	0.33	0.09
E027	word	glass	0.52	glæs	g, l, s	æ	g→l, l→æ æ→s		0.09	0.06	0.33	0.08
E028	word	bread	0.52	bɹɛd	b, ɹ, d	ɛ	b→ɹ, ɹ→ɛ ɛ→d		0.09	0.06	0.33	0.08
E029	word	cheese	0.48	tʃiːz	tʃ, z	iː	tʃ→iː, iː→z		0.07	0.06	0.33	0.07
E030	word	fruit	0.52	fɹuːt	f, ɹ, t	uː	f→ɹ, ɹ→uː uː→t		0.09	0.06	0.33	0.08
E031	word	milk	0.52	mɪlk	m, l, k	ɪ	m→ɪ, ɪ→l l→k	dark l	0.09	0.06	0.33	0.09
E032	word	egg	0.48	ɛg	g	ɛ	ɛ→g		0.05	0.06	0.33	0.06
E033	word	orange	0.80	ɔɹɪndʒ	n, dʒ	ɔɹ, ɪ	ɔɹ→ɪ, ɪ→n n→dʒ	r coloring	0.09	0.06	0.67	0.15
E034	word	banana	1.04	bənænə	b, n	ə, æ	b→ə, ə→n n→æ, æ→n n→ə	vowel reduction	0.09	0.00	0.67	0.14
E035	word	tomato	1.04	təmeɪroʊ	t, m	ə, eɪ, oʊ	t→ə, ə→m m→eɪ, eɪ→t t→oʊ	vowel reduction, flapping	0.12	0.00	0.67	0.16
E036	word	potato	1.04	pəteɪroʊ	p, t	ə, eɪ, oʊ	p→ə, ə→t t→eɪ, eɪ→t t→oʊ	vowel reduction, flapping	0.12	0.00	0.67	0.16
E037	word	coffee	0.76	kɔfi	k, f	ɔ, i	k→ɔ, ɔ→f f→i		0.09	0.00	0.67	0.11
E038	word	teacher	0.76	tiːtʃɝ	t, tʃ	iː, ɝ	t→iː, iː→tʃ tʃ→ɝ	r coloring	0.09	0.00	0.67	0.14
E039	word	student	0.80	stuːdənt	s, t, d, n	uː, ə	s→t, t→uː uː→d, d→ə ə→n, n→t	vowel reduction	0.14	0.06	0.67	0.16
E040	word	doctor	0.80	dɔktɝ	d, k, t	ɔ, ɝ	d→ɔ, ɔ→k k→t, t→ɝ	r coloring	0.12	0.00	0.67	0.15

Continued on next page

ID	Type	Text	Dur.	Pron.	Cons.	Nuc.	Tr	Ph	C_{ph}	C_{coda}	C_{stress}	U_A
E041	word	singer	0.76	sɪŋɚ	s, ŋ	ɪ, ɚ	s→ɪ, ɪ→ŋ ŋ→ɚ	r coloring	0.09	0.00	0.67	0.14
E042	word	music	0.80	mjuːzɪk	m, j, z, k	uː, ɪ	m→j, j→uː uː→z, z→ɪ ɪ→k		0.14	0.06	0.67	0.13
E043	word	paper	0.76	peɪpɚ	p	eɪ, ɚ	p→eɪ, eɪ→p p→ɚ	aspiration; r coloring	0.07	0.00	0.67	0.15
E044	word	pencil	0.80	pɛnsəl	p, n, s, l	ɛ, ə	p→ɛ, ɛ→n n→s, s→ə ə→l	vowel reduction, dark l	0.14	0.06	0.67	0.18
E045	word	window	0.80	wɪndoʊ	w, n, d	ɪ, oʊ	w→ɪ, ɪ→n n→d, d→oʊ		0.12	0.00	0.67	0.12
E046	word	garden	0.76	gɑɹdən	g, d, n	ɑɹ, ə	g→ɑɹ, ɑɹ→d d→ə, ə→n	vowel reduction, r coloring	0.12	0.06	0.67	0.18
E047	word	button	0.76	bʊʔən	b, t, n	ʊ, ə	b→ʊ, ʊ→t t→ə, ə→n	glottalization, vowel reduction	0.12	0.06	0.67	0.17
E048	word	little	0.76	lɪɾəl	l, t	ɪ, ə	l→ɪ, ɪ→t t→ə, ə→l	flapping, vowel reduction, dark l	0.09	0.06	0.67	0.18
E049	word	city	0.76	sɪɾi	s, t	ɪ, i	s→ɪ, ɪ→t t→i	flapping	0.09	0.00	0.67	0.13
E050	word	better	0.76	bɛɾɚ	b, t	ɛ, ɚ	b→ɛ, ɛ→t t→ɚ	flapping, r coloring	0.09	0.00	0.67	0.16
E051	word	writer	0.76	ɹaɪɾɚ	ɹ, t	aɪ, ɚ	ɹ→aɪ, aɪ→t t→ɚ	flapping, r coloring	0.09	0.00	0.67	0.16
E052	word	winter	0.80	wɪntɚ	w, n, t	ɪ, ɚ	w→ɪ, ɪ→n n→t, t→ɚ	r coloring	0.12	0.00	0.67	0.15
E053	word	spring	0.52	spɹɪŋ	s, p, ɹ, ŋ	ɪ	s→p, p→ɹ ɹ→ɪ, ɪ→ŋ		0.12	0.06	0.33	0.08
E054	word	street	0.52	stɹiːt	s, t, ɹ	iː	s→t, t→ɹ ɹ→iː, iː→t		0.09	0.06	0.33	0.08
E055	word	strong	0.52	stɹɔŋ	s, t, ɹ, ŋ	ɔ	s→t, t→ɹ ɹ→ɔ, ɔ→ŋ		0.12	0.06	0.33	0.08
E056	word	world	0.52	wɝld	w, l, d	ɝ	w→ɝ, ɝ→l l→d	r coloring, dark l	0.09	0.06	0.33	0.12
E057	word	help	0.52	hɛlp	h, l, p	ɛ	h→ɛ, ɛ→l l→p	dark l	0.09	0.06	0.33	0.09
E058	word	desk	0.52	dɛsk	d, s, k	ɛ	d→ɛ, ɛ→s s→k		0.09	0.06	0.33	0.08
E059	word	next	0.52	nɛkst	n, k, s, t	ɛ	n→ɛ, ɛ→k k→s, s→t		0.12	0.06	0.33	0.08
E060	word	asked	0.52	æskt	s, k, t	æ	æ→s, s→k k→t		0.09	0.06	0.33	0.08
E061	word	sixth	0.52	sɪksθ	s, k, θ	ɪ	s→ɪ, ɪ→k k→s, s→θ		0.09	0.06	0.33	0.08
E062	word	texts	0.52	tɛksts	t, k, s	ɛ	t→ɛ, ɛ→k k→s, s→t t→s		0.09	0.06	0.33	0.08
E063	word	blue	0.52	bluː	b, l	uː	b→l, l→uː		0.07	0.00	0.33	0.06
E064	word	green	0.52	gɹiːn	g, ɹ, n	iː	g→ɹ, ɹ→iː iː→n		0.09	0.06	0.33	0.08
E065	word	brown	0.52	bɹaʊn	b, ɹ, n	aʊ	b→ɹ, ɹ→aʊ aʊ→n		0.09	0.06	0.33	0.08
E066	word	quick	0.52	kwɪk	k, w	ɪ	k→w, w→ɪ ɪ→k		0.07	0.06	0.33	0.07
E067	word	quiet	0.80	kwaɪət	k, w, t	aɪ, ə	k→w, w→aɪ aɪ→ə, ə→t	vowel reduction	0.12	0.06	0.67	0.16
E068	word	around	0.80	əɹaʊnd	ɹ, n, d	ə, aʊ	ə→ɹ, ɹ→aʊ aʊ→n, n→d	vowel reduction	0.12	0.06	0.67	0.16
E069	word	away	0.76	əweɪ	w	ə, eɪ	ə→w, w→eɪ	vowel reduction	0.07	0.00	0.67	0.14
E070	word	about	0.76	əbaʊt	b, t	ə, aʊ	ə→b, b→aʊ aʊ→t	vowel reduction	0.09	0.06	0.67	0.15
E071	word	today	0.76	tədeɪ	t, d	ə, eɪ	t→ə, ə→d d→eɪ	vowel reduction	0.09	0.00	0.67	0.14
E072	word	again	0.76	əgɛn	g, n	ə, ɛ	ə→g, g→ɛ ɛ→n	vowel reduction	0.09	0.06	0.67	0.15
E073	word	machine	0.76	məʃiːn	m, ʃ, n	ə, iː	m→ə, ə→ʃ ʃ→iː, iː→n	vowel reduction	0.12	0.06	0.67	0.15
E074	word	photograph	1.08	foʊɾəgɹæf	f, t, g, ɹ	oʊ, ə, æ	f→oʊ, oʊ→t t→ə, ə→g g→ɹ, ɹ→æ æ→f	flapping, vowel reduction	0.16	0.06	0.67	0.18
E075	word	photography	1.36	fətɔgɹəfi	f, t, g, ɹ	ə, ɔ, i	f→ə, ə→t t→ɔ, ɔ→g g→ɪ, ɹ→ə ə→f, f→i	vowel reduction	0.16	0.00	0.67	0.16

Continued on next page

ID	Type	Text	Dur.	Pron.	Cons.	Nuc.	Tr	Ph	C_{ph}	C_{coda}	C_{stress}	U_A
E076	word	electricity	1.64	ɪlɛktɹɪsəri	l, k, t, ɹ, s	ɪ, ɛ, ə, i	ɪ→l, l→ɛ ɛ→k, k→t t→ɹ, ɹ→ɪ ɪ→s, s→ə ə→t, t→i	flapping, vowel reduction	0.21	0.00	0.67	0.18
E077	phrase	pick it up	1.10	pɪk ɪt ʊp	p, k, t	ɪ, ʊ	p→ɪ, ɪ→k ɪ→t, ʊ→p	linking	0.12	0.19	0.50	0.13
E078	phrase	turn it on	1.10	tɝn ɪrɑn	t, n	ɝ, ɪ, ɑ	t→ɝ, ɝ→n ɪ→t, ɑ→n	linking, flapping, r coloring	0.12	0.12	0.50	0.17
E079	phrase	take a seat	1.10	teɪk əsiːt	t, k, s	eɪ, ə, iː	t→eɪ, eɪ→k s→iː, iː→t	linking, weak form reduction	0.14	0.12	0.83	0.22
E080	phrase	read it aloud	1.66	ɹiːd ɪrəlaʊd	ɹ, d, t, l	iː, ɪ, ə, aʊ	ɹ→iː, iː→d ɪ→t, ə→l l→aʊ, aʊ→d	linking, flapping, vowel reduction	0.19	0.12	0.83	0.23
E081	phrase	put it away	1.66	pʊɾɪɾəweɪ	p, t, w	ʊ, ɪ, ə, eɪ	p→ʊ, ʊ→t ɪ→t, ə→w w→eɪ	linking, flapping, vowel reduction	0.16	0.06	0.83	0.22
E082	phrase	get a ticket	1.66	gɛɾətɪkɪt	g, t, k	ɛ, ə, ɪ	g→ɛ, ɛ→t t→ɪ, ɪ→k k→ɪ, ɪ→t	flapping, weak form reduction	0.14	0.06	0.83	0.21
E083	phrase	a cup of tea	1.66	əkʊp əv tiː	k, p, v, t	ə, ʊ, iː	k→ʊ, ʊ→p ə→v, t→iː	weak form reduction, linking	0.16	0.12	0.83	0.24
E084	phrase	a glass of water	1.98	əglæs əv wɔɾɚ	g, l, s, v, w, t	ə, æ, ɔ, ɚ	g→l, l→æ æ→s, ə→v w→ɔ, ɔ→t t→ɚ	weak form reduction, flapping, r coloring	0.23	0.12	0.83	0.29
E085	phrase	an apple a day	1.94	ən æpəl ədeɪ	n, p, l, d	ə, æ, eɪ	ə→n, æ→p p→ə, ə→l d→eɪ	linking, weak form reduction, dark l	0.16	0.12	0.83	0.24
E086	phrase	black coffee please	1.74	blæk kɔfi pliːz	b, l, k, f, p, z	æ, ɔ, i, iː	b→l, l→æ æ→k, k→ɔ ɔ→f, f→i p→l, l→iː iː→z		0.23	0.12	0.83	0.19
E087	phrase	fresh fruit salad	1.70	fɹɛʃfɹuːt sæləd	f, ɹ, ʃ, t, s, l, d	ɛ, uː, æ, ə	f→ɹ, ɹ→ɛ ɛ→ʃ, f→ɹ ɹ→uː, uː→t s→æ, æ→l l→ə, ə→d	vowel reduction, dark l	0.26	0.19	0.83	0.24
E088	phrase	good morning	1.10	gʊd mɔɹnɪŋ	g, d, m, n, ŋ	ʊ, ɔɹ, ɪ	g→ʊ, ʊ→d m→ɔɹ, ɔɹ→n n→ɪ, ɪ→ŋ n→aɪ, aɪ→s	r coloring	0.19	0.12	0.83	0.20
E089	phrase	nice to meet you	1.66	naɪs təmiːt juː	n, s, t, m, j	aɪ, ə, iː, uː	t→ə, m→iː iː→t, j→uː d→ɪ, ɪ→d	weak form reduction	0.21	0.12	0.83	0.22
E090	phrase	did you call	1.38	dɪdʒəkɔl	d, j, k, l	ɪ, ə, ɔ	j→ə, k→ɔ ɔ→l	yod coalescence, weak form reduction, dark l	0.16	0.12	0.83	0.23
E091	phrase	want to go	1.38	wɑnəgoʊ	w, n, t, g	ɑ, ə, oʊ	w→ɑ, ɑ→n n→t, t→ə g→oʊ	weak form reduction	0.16	0.06	0.83	0.20
E092	phrase	going to work	1.38	goʊɪŋtəwɝk	g, ŋ, t, w, k	oʊ, ɪ, ə, ɝ	g→oʊ, oʊ→ɪ ɪ→ŋ, t→ə w→ɝ, ɝ→k	weak form reduction, r coloring	0.21	0.12	0.83	0.25
E093	phrase	could have known	1.38	kʊd əv noʊn	k, d, v, n	ʊ, ə, oʊ	k→ʊ, ʊ→d ə→v, n→oʊ oʊ→n	weak form reduction, linking	0.16	0.19	0.83	0.23
E094	phrase	next train home	1.46	nɛkst tɪem hoʊm	n, k, s, t, ɹ, h, m	ɛ, eɪ, oʊ	n→ɛ, ɛ→k k→s, s→t t→ɹ, ɹ→eɪ eɪ→n, h→oʊ oʊ→m		0.23	0.19	0.50	0.15
E095	phrase	world class service	1.46	wɝld klæs sɝvɪs	w, l, d, k, s, v	ɝ, æ, ɪ	w→ɝ, ɝ→l l→d, k→l l→æ, æ→s s→ɝ, ɝ→v v→ɪ, ɪ→s s→ɪ, ɪ→k	r coloring, dark l	0.21	0.12	0.83	0.23
E096	phrase	six small streets	1.46	sɪks smɔl stɹiːts	s, k, m, l, t, ɹ	ɪ, ɔ, iː	k→s, s→m m→ɔ, ɔ→l s→t, t→ɹ ɹ→iː, iː→t t→s	dark l	0.21	0.12	0.50	0.15

E Discussion and Limitations

LEAP reframes few-shot voice personalization as an active enrollment problem: under a fixed recording-time budget, the value of an utterance depends on the linguistic evidence already collected. Across matched-budget evaluation, held-out speakers, budgets, lexical realizations, languages, and voice-personalization models, the results show that modeling this set-conditioned marginal value improves both automatic utility and listener preference. Our zero-shot language-transfer setting eliminates target-language downstream utility outcomes, policy retraining, and model selection. It nevertheless assumes an eligible target-language prompt inventory and text-side linguistic descriptors. These inputs specify the available enrollment interface—the sentences a user may be asked to record and the evidence they contain—rather than requiring new target-language personalization experiments. Developing automatic or weakly supervised ways to instantiate these interfaces from lexical and phonological resources is a promising direction for extending the approach to lower-resource languages. LEAP selects enrollment content from text-side evidence and estimated recording cost. It does not explicitly optimize recording-environment factors such as microphone quality, background noise, or deviations between prompted and realized speech. These factors are complementary to content selection and could be incorporated through an acoustic quality estimate or an adaptive post-recording policy. Finally, although we evaluate multiple languages and voice-personalization backends, broader coverage of language families, speaking styles, and personalization paradigms remains important future work.