

Auditing Speaker-Level Metadata Utility in Text-Residual Prosody Prediction

Gaeun Kim, Jihyun Park, Sohee Jang, Joon-Hyuk Chang

Hanyang University, Seoul, Republic of Korea

{gaeun8389, jjhh0210, shjang97, jchang}@hanyang.ac.kr

Abstract

Text only partially determines prosodic realization, leaving residual variation that may align with speaker-level metadata such as regional accent or first-language (L1) background. However, because these labels are assigned to speakers, their predictive gains may overlap with broader speaker-associated structure rather than uniquely reflect category-level prosodic effects. We frame this as an attribution problem and propose a speaker-controlled audit using matched text-residual prosody predictors that differ only in access to metadata labels and speaker identity. The audit measures label utility beyond text, speaker utility beyond text, and the additional empirical utility of the label when speaker conditioning is already available. Across VCTK regional-accent labels and L2-ARCTIC L1 labels, metadata yields modest gains over text, while speaker conditioning yields larger gains and leaves little additional label utility under speaker control. Random speaker-group and label-perturbation controls further characterize the specificity and use of the supplied metadata. Importantly, speaker-held-out evaluation shows that metadata transfer to unseen speakers is limited and target-dependent, indicating that observed-speaker metadata gains do not by themselves establish speaker-independent category-level effects. These results position speaker-controlled auditing as a prerequisite attribution check before assigning category-level interpretations to metadata-conditioned predictive gains.

1 Introduction

The same words can be realized with different prosodic patterns in speech. Prosody—including prominence, phrasing, timing, and pitch movement—signals information structure and shapes how utterances are interpreted (Jun, 2005; Ladd, 2008; Wagner and Watson, 2010; Cole, 2015). Text provides important cues to prosody, with prior work linking linguistic predictability and

information structure to duration, prominence, and other prosodic realizations (Aylett and Turk, 2006; Seyfarth, 2014; Breen et al., 2010; Malisz et al., 2018). Text-based prosody prediction therefore provides a useful setting for studying both text-recoverable structure and residual variation associated with non-textual factors (Talman et al., 2019; Suni et al., 2020; Wolf et al., 2023). We use automatically extracted acoustic prosody targets as diagnostic measurements of this residual variation rather than as perceptual ground truth.

Predictive gains from non-textual metadata, however, are not self-interpreting. In the corpora considered here, regional-accent and L1 labels are assigned at the speaker level, with each category represented by a limited number of speakers. Their predictive utility may therefore overlap with broader speaker-associated structure rather than uniquely reflect category-level prosodic effects. We use speaker identity as a conservative control for such structure, which may include individual realization habits, pronunciation, speaking style, recording conditions, and the metadata category itself. Accordingly, a small label gain after speaker control does not imply the absence of category-level effects; it indicates that the observed gain alone is insufficient to attribute the improvement specifically to the metadata category.

We formulate this as a speaker-controlled attribution audit for text-residual prosody prediction. Matched predictors differ only in access to metadata labels and speaker identity, allowing us to measure label utility beyond text, speaker utility beyond text, and additional empirical label utility under speaker control. Because the label is determined by speaker identity in these corpora, the last quantity is an empirical held-out error contrast between matched predictors. Our contribution is the matched audit of how metadata-conditioned utility changes under speaker control. We complement this decomposition with likelihood-based scoring

and diagnostics for label specificity, dependence, and residual structure.

We instantiate the audit with regional-accent labels in VCTK (Veaux et al., 2013) and L1 labels in L2-ARCTIC (Zhao et al., 2018). For observed speakers, metadata yields modest gains beyond text, while speaker conditioning yields larger gains and leaves little additional label utility under speaker control. Speaker-held-out evaluation addresses the distinct question of transfer to unseen speakers and shows limited, target-dependent effects. Together, these results show that observed-speaker metadata gains do not by themselves establish speaker-independent category-level effects and motivate speaker-controlled auditing before stronger category-level interpretations.

2 Related Work

2.1 Text-Based Prosody Prediction

Text-based prosody prediction has been studied both for speech generation and as a diagnostic of how prosodic realization relates to linguistic context. Duration, pitch, energy, and related representations are commonly modeled in speech generation (Ren et al., 2021; Li et al., 2023b), while diagnostic studies predict prominence, boundaries, and other acoustic cues from text (Talman et al., 2019; Suni et al., 2020; Dhamne et al., 2025) or quantify text–prosody redundancy directly (Wolf et al., 2023). We instead ask how residual predictive gains should be attributed when non-textual metadata is assigned at the speaker level.

2.2 Speaker-Level Prosodic Variation

Prosodic realization varies across speakers, regional varieties, and linguistic backgrounds. Prior work documents regional and dialectal variation in intonation, pitch-accent realization, and prominence (Grabe et al., 2000; Grabe and Post, 2002; Clopper and Smiljanic, 2011; Smith and Rathcke, 2020). Such findings motivate speaker-level metadata as a conditioning source, but a predictive gain need not uniquely reflect a category-level effect when the label is structurally coupled with speaker identity. Our analysis therefore treats speaker identity as a conservative control for broader speaker-associated structure.

2.3 Attribution under Speaker-Level Metadata

Our analysis is related to variance-partitioning and probing approaches that characterize how measured variation is associated with different sources of information. We operationalize attribution for text-residual prosody when metadata is structurally attached to speakers. Matched text-only, text+label, text+speaker, and full predictors quantify how metadata utility changes under speaker control, complemented by random-group, label-perturbation, and residual diagnostics. Our approach provides a speaker-controlled operationalization of attribution analysis for speaker-level metadata.

3 Method

3.1 Diagnostic Prediction Setup

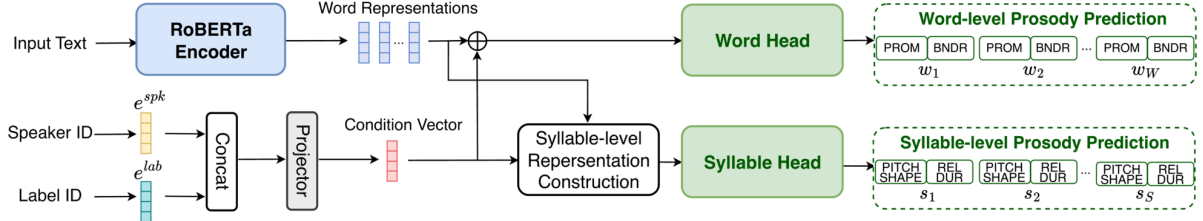
We use text-based prosody prediction as a diagnostic instrument for metadata attribution. The predictor comparisons are designed to interpret conditioning gains, not to select the best conditioning model.

Let x denote the input text, ℓ the available speaker-level metadata label, s the speaker identity, and $\mathbf{Y}$ the acoustically derived prosody target. For VCTK, ℓ is the regional-accent label; for L2-ARCTIC, ℓ is the L1 label. In both corpora, ℓ is assigned at the speaker level, so ℓ and s are structurally entangled. We therefore compare matched predictors that share the same text input, architecture, target extraction pipeline, data splits, and optimization setup, and differ only in access to ℓ , s , both, or neither. This provides matched attribution contrasts for label utility beyond text, speaker utility beyond text, and the additional empirical utility of label conditioning when speaker conditioning is already available.

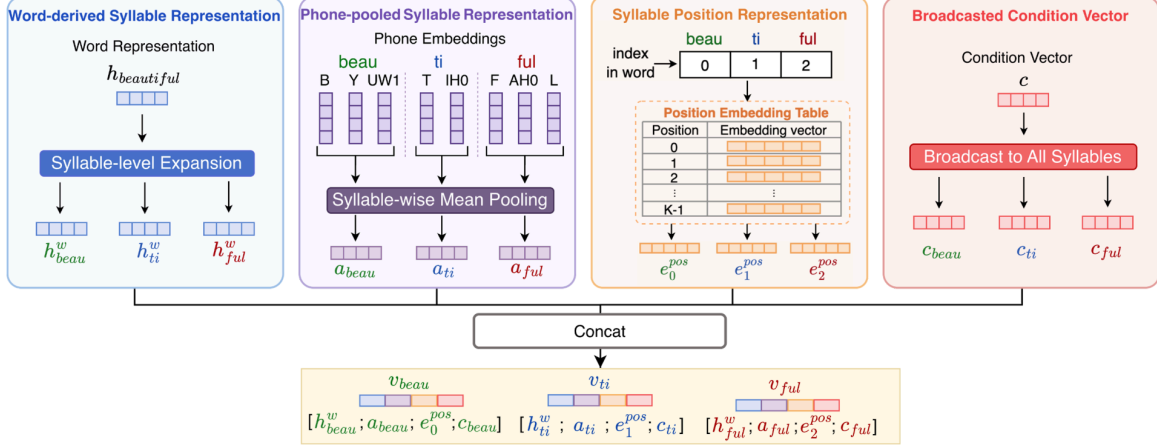
Hierarchical prosody targets. We represent prosody with word- and syllable-level targets:

$$\mathbf{Y} = \{\mathbf{Y}_{1:W}^{\text{word}}, \mathbf{Y}_{1:S}^{\text{syll}}\}, \quad (1)$$

where W and S are the numbers of words and syllables in an utterance. Concretely, $\mathbf{Y}_j^{\text{word}} = [\bar{m}_j, \bar{b}_j]$ contains word-level prominence and boundary strength, which capture emphasis and phrasing. The syllable-level target $\mathbf{Y}_i^{\text{syll}} = [p_{i,1}, p_{i,2}, p_{i,3}, r_i]$ contains a three-point pitch-shape contour and relative duration, capturing local intonational movement and timing. These quantities are defined in



(a) Matched prosody prediction model.



(b) Syllable-level representation construction.

Figure 1: Architecture of the matched prosody predictors. (a) The predictor encodes text with RoBERTa, combines optional speaker and metadata-label embeddings into a condition vector, and uses this conditioning for word- and syllable-level prosody prediction. (b) Syllable-level representations are formed by concatenating word-derived syllable representations, phone-pooled syllable representations, syllable-position embeddings, and the broadcasted condition vector. In (b), h_i^w , a_i , e_i^{pos} , and c_i denote these four components for syllable i , and v_i denotes the final concatenated syllable representation. Only the available conditioning variables change across settings.

Appendix A.4. Prominence and boundary strength are extracted from wavelet-based prosody representations (Sun et al., 2017) and robustly normalized using train-split statistics. Pitch shape is represented as an utterance-normalized three-point semitone F_0 contour, and relative duration is normalized within each utterance to reduce sensitivity to global speaking rate.

These four targets were selected to span complementary prosodic dimensions—emphasis, phrasing, local intonational movement, and timing—rather than to exhaust accent- or L1-related prosodic variation. Other properties, including speaking rate, pause duration or frequency, pitch range or variability, and energy contours, remain outside the present target representation. The automatically extracted targets are used as diagnostic acoustic measurements rather than as perceptual ground truth.

Matched predictors. Figure 1 shows the shared predictor architecture used in the matched com-

parisons. For each conditioning set C , we train a predictor

$$\hat{\mathbf{Y}}_C = f_\theta(x; C), \quad (2)$$

with

$$C \in \{\emptyset, \{\ell\}, \{s\}, \{\ell, s\}\}. \quad (3)$$

These correspond to text-only, text+label, text+speaker, and full predictors; the full model uses both label and speaker identity. All main regression predictors share the same RoBERTa encoder (Liu et al., 2019), pooling and representation construction, target extraction, masks, and training/evaluation protocol. They differ only in access to the speaker-level metadata label and/or speaker identity. When available, the metadata-label and speaker-identity embeddings are concatenated and projected to the encoder hidden dimension to form a learned condition vector. Word-level predictions are made from conditioned word representations; syllable-level predictions concatenate syllable-aligned word

representations, phone-pooled syllable representations, syllable-position embeddings, and the broadcasted condition vector, as detailed in Figure 1(b). This construction preserves contextual lexical information, syllable-local phonological form, and within-word syllable position as distinct components before predicting syllable-level pitch shape and duration. The main prediction heads are trained with masked MSE regression losses over valid aligned units. The likelihood-based diagnostic uses a separate otherwise matched Gaussian NLL variant that keeps the same mean prediction heads and adds parallel log-variance heads. Additional architecture and optimization details are given in Appendix A.1 and Appendix A.2.

3.2 Residual Utility Decomposition

We quantify metadata utility with two matched residual-utility views. The first uses MAE reductions to measure point-prediction gains, and the second uses held-out Gaussian NLL reductions as a complementary likelihood-based diagnostic.

MAE-based residual utility. The text-only model provides an operational estimate of the component of prosody predictable from the available text representation. We then measure how much speaker-level metadata labels and speaker identity reduce the remaining held-out prediction error. For target-wise notation, let $y^{(t)}$ denote the component of $\mathbf{Y}$ corresponding to target type t . For target t , let $E_t(C)$ denote held-out MAE under conditioning set C , computed over valid aligned units. We define MAE-based residual utility as:

$$\begin{aligned} U_t(\ell) &= E_t(\emptyset) - E_t(\{\ell\}), \\ U_t(s) &= E_t(\emptyset) - E_t(\{s\}), \\ U_t(\ell | s) &= E_t(\{s\}) - E_t(\{\ell, s\}). \end{aligned} \quad (4)$$

Here, $U_t(\ell)$ measures speaker-level label utility beyond text, $U_t(s)$ measures speaker utility beyond text, and $U_t(\ell | s)$ measures the additional empirical utility of explicit label conditioning when speaker identity is already available. Positive values indicate lower held-out MAE relative to the corresponding baseline.

Because ℓ is determined by s in these corpora, $U_t(\ell | s)$ is interpreted as an operational held-out error contrast between separately trained predictors, rather than as speaker-independent category information.

We use speaker identity as a conservative control for broader speaker-associated variation rather than

as a causal explanatory factor. The comparison is intentionally asymmetric because speaker identity is the finer-grained unit that carries the metadata label and may capture speaker-specific characteristics and recording-related variation. The relevant attribution question is therefore how much of an apparent metadata gain survives this broader speaker control.

A large $U_t(\ell)$ with a small $U_t(\ell | s)$ indicates that the speaker-level label is predictive beyond text, but that its utility largely overlaps with speaker-associated structure. A positive $U_t(\ell | s)$ indicates that explicit label conditioning further reduces held-out error within the matched predictor family after speaker identity is already provided. These quantities are predictive, not causal: they characterize the utility of the observed speaker-level metadata under the corpus labeling structure rather than causal accent- or L1-related prosodic effects.

Likelihood-based diagnostic utility. The MAE-based utilities measure reductions in point-prediction error. As a complementary probabilistic diagnostic, we train otherwise matched Gaussian NLL variants and compute target-wise held-out negative log-likelihood over valid target units using the same masks as the regression objectives. The NLL variant keeps the same mean prediction heads as the regression model and adds parallel heads that predict a diagonal log-variance for each target dimension; the constant term in the Gaussian NLL is omitted because it is identical across matched models.

For compactness, let $Z_t(C)$ denote the held-out Gaussian NLL per valid target unit for target group t under conditioning set C :

$$Z_t(C) = \frac{1}{N_t} \sum_{(n,k) \in \mathcal{D}_t} -\log q_\theta(y_{n,k}^{(t)} | X_n, C_n). \quad (5)$$

Here, t denotes prominence, boundary strength, pitch shape, or relative duration; n indexes the input utterance; k indexes a valid target unit; C_n is one of $\emptyset$, $\{\ell\}$, $\{s\}$, or $\{\ell, s\}$; $y_{n,k}^{(t)}$ is the target value; and $\mathcal{D}_t$ contains the valid masked units for target group t .

We then compute target-wise likelihood-based

utilities:

$$\begin{aligned}\hat{I}_{\text{NLL}}^{(t)}(\ell | X) &= Z_t(\emptyset) - Z_t(\{\ell\}), \\ \hat{I}_{\text{NLL}}^{(t)}(s | X) &= Z_t(\emptyset) - Z_t(\{s\}), \\ \hat{I}_{\text{NLL}}^{(t)}(\ell | X, s) &= Z_t(\{s\}) - Z_t(\{\ell, s\}).\end{aligned}\quad (6)$$

Positive values indicate that the added conditioning variable lowers held-out conditional cross-entropy under the same matched contrasts used for MAE-based utility. As with the MAE-based quantities, these are operational likelihood contrasts within the matched predictor family rather than estimates of causal or speaker-independent category information. We use them as complementary diagnostics of the conditioning order.

3.3 Diagnostics for Interpreting Metadata Utility

We organize the audit around five diagnostics. The first two address the attribution contrasts defined above; the remaining diagnostics examine label specificity, label dependence, residual structure, and acoustic realization.

Q1: Does the label reduce residual error beyond text? The text-only versus text+label contrast measures $U_t(\ell)$, and the likelihood-based diagnostic evaluates the same contrast using held-out Gaussian NLL.

Q2: How much label utility remains after speaker control? The text+speaker versus full contrast measures $U_t(\ell | s)$. This is an observed-speaker attribution contrast. We separately evaluate speaker-held-out prediction as an unseen-speaker transfer diagnostic, asking whether category-associated regularities learned from training speakers generalize to speakers not seen during training.

Q3: Are the gains specific to the observed label structure? Random speaker-group controls test whether metadata gains exceed arbitrary same-cardinality speaker partitions. Prompt-matched residual proximity and residual-variance analysis compare label- and speaker-associated residual structure. These diagnostics assess specificity to the observed grouping but do not by themselves establish a causal or purely linguistic category effect.

Q4: Do trained predictors use the supplied label input? Shuffled-label evaluation and label-swap sensitivity perturb fixed trained predictors at evaluation time. These diagnostics test dependence

on the supplied label input rather than additional utility after speaker control.

Q5: Are the predicted plans acoustically active? A StyleTTS2 (Li et al., 2023b) realization check tests whether injected predicted plans induce measurable pitch, energy, and duration differences.

4 Results

4.1 Experimental Setup

We evaluated the audit on aligned subsets of VCTK (Veaux et al., 2013) and L2-ARCTIC (Zhao et al., 2018) with available word- and syllable-level prosody targets. VCTK provides speaker-level regional-accent labels, while L2-ARCTIC provides speaker-level L1 labels. The main VCTK split contains 107 speakers across 11 regional-accent categories, with substantial imbalance in speakers per category; L2-ARCTIC contains 24 speakers across six L1 categories, with four speakers per L1. Detailed speaker-label distributions and split statistics are reported in Appendix A.3.

The main experiments used shared utterance-level train, validation, and test splits across all conditioning settings and seeds. The same speaker pool therefore appears across these splits. This observed-speaker regime is the primary attribution setting because it directly tests how apparent metadata utility changes when the individual speaker carrying that metadata is also modeled. We separately evaluate speaker-held-out splits to ask whether metadata-associated regularities generalize to speakers not seen during either training or validation.

We use the same text-conditioned predictor across all matched conditions so that text-predictable prosodic structure is controlled consistently when measuring the additional utility of speaker or metadata conditioning. All matched conditions share the same text encoder, architecture, target extraction, data splits, optimization, and evaluation protocol, so the quantities of interest are the relative conditioning contrasts rather than absolute model performance.

Checkpoints were selected on the validation split and evaluated once on the fixed test split. Main prediction and residual-utility results were averaged over five fixed random seeds. We report masked MAE and Pearson correlation for each target type, computed only over valid word or syllable positions under the corresponding target masks. For residual-utility measures, positive values indicate

Model	Prom.		Boundary		PitchShape		RelDur	
	MAE↓	Corr.↑	MAE↓	Corr.↑	MAE↓	Corr.↑	MAE↓	Corr.↑
VCTK								
Text-only	0.327±0.001	0.761±0.001	0.301±0.002	0.699±0.001	1.445±0.009	0.559±0.0004	0.180±0.001	0.902±0.001
Text+Acc.	0.322±0.001	0.767±0.001	0.298±0.001	0.704±0.001	1.433±0.003	0.567±0.001	0.179±0.001	0.904±0.002
Text+Speaker	0.304±0.002	0.792±0.002	0.294±0.001	0.717±0.001	1.380±0.004	0.601±0.001	0.177±0.001	0.907±0.001
Text+Acc.+Spk.	0.303±0.002	0.793±0.002	0.292±0.001	0.719±0.001	1.377±0.003	0.604±0.001	0.175±0.002	0.908±0.001
L2-ARCTIC								
Text-only	0.413±0.001	0.656±0.001	0.373±0.001	0.638±0.001	1.556±0.002	0.549±0.001	0.228±0.001	0.832±0.001
Text+L1	0.392±0.001	0.691±0.001	0.363±0.001	0.659±0.001	1.499±0.002	0.584±0.001	0.222±0.001	0.840±0.001
Text+Speaker	0.377±0.001	0.716±0.001	0.351±0.001	0.684±0.001	1.415±0.001	0.629±0.001	0.219±0.001	0.845±0.001
Text+L1+Spk.	0.376±0.0008	0.718±0.0013	0.351±0.0009	0.686±0.0007	1.410±0.0005	0.632±0.0007	0.218±0.0005	0.847±0.0007

Table 1: Main prosody prediction results in the observed-speaker regime. Values are mean $\pm$ standard deviation over five fixed random seeds. VCTK uses regional-accent labels; L2-ARCTIC uses L1 labels.

Target	$U_t(\ell)$	$U_t(s)$	$U_t(\ell s)$
VCTK			
Prom.	0.0050±0.0008	0.0229±0.0026	0.0009±0.0005
Boundary	0.0031±0.0019	0.0077±0.0018	0.0012±0.0005
PitchShape	0.0122±0.0095	0.0659±0.0086	0.0027±0.0045
RelDur	0.0010±0.0020	0.0037±0.0019	0.0015±0.0014
L2-ARCTIC			
Prom.	0.0205±0.0009	0.0353±0.0005	0.0017±0.0007
Boundary	0.0101±0.0005	0.0221±0.0014	0.0006±0.0004
PitchShape	0.0568±0.0009	0.1414±0.0018	0.0044±0.0007
RelDur	0.0055±0.0003	0.0091±0.0007	0.0008±0.0003

Table 2: Main residual-utility decomposition based on held-out MAE reductions. Here, ℓ denotes regional accent for VCTK and L1 for L2-ARCTIC. Values are mean $\pm$ standard deviation over five fixed random seeds. Seed-level positivity counts and paired bootstrap intervals are reported in Appendix B.1.

held-out error reduction relative to the corresponding baseline. Implementation, split, optimization, and diagnostic details are provided in Appendix A.

4.2 Main Decomposition: Metadata Utility and Speaker Control

Main matched prediction results. Table 1 reports matched test-set prediction results in the observed-speaker regime. The first audit question is whether speaker-level metadata reduces residual prediction error beyond text, and whether that gain remains when speaker identity is available. We therefore focus on paired differences across conditioning settings and seed consistency rather than raw cross-corpus comparisons.

Label-conditioned gains were strongest on L2-ARCTIC. Adding L1 labels improved over text-only prediction for all target groups, with especially consistent gains for prominence and pitch shape.

Target	$\hat{I}_{\text{NLL}}(\ell X)$	$\hat{I}_{\text{NLL}}(s X)$	$\hat{I}_{\text{NLL}}(\ell X, s)$
VCTK			
Prom.	0.0138±0.0049	0.0602±0.0063	0.0083±0.0123
Boundary	0.0128±0.0144	0.0227±0.0143	-0.0038±0.0091
PitchShape	0.0127±0.0044	0.1211±0.0056	0.0015±0.0015
RelDur	0.0101±0.0027	0.0385±0.0042	0.0033±0.0023
L2-ARCTIC			
Prom.	0.0368±0.0036	0.0636±0.0054	0.0066±0.0009
Boundary	0.0108±0.0040	0.0200±0.0183	0.0090±0.0071
PitchShape	0.0333±0.0019	0.0938±0.0016	0.0034±0.0006
RelDur	0.0312±0.0030	0.0537±0.0027	0.0038±0.0010

Table 3: Target-wise likelihood-based utility from matched Gaussian predictors. Values are reductions in held-out Gaussian NLL, measured in nats per valid target unit and reported as mean $\pm$ standard deviation over five fixed random seeds. Positive values indicate lower held-out conditional cross-entropy under the added conditioning variable. Here, ℓ denotes regional accent for VCTK and L1 for L2-ARCTIC.

However, speaker conditioning yielded larger gains than L1 conditioning, and the full model added only small improvements over the speaker-conditioned model. VCTK showed the same ordering with weaker and less stable regional-accent gains: accent labels improved over text-only prediction on average, but speaker conditioning yielded substantially larger gains. Thus, the matched results support the central attribution pattern: metadata labels provided residual signal beyond text, but explicit speaker identity captured substantially more of the remaining variation.

MAE-based residual utility. Table 2 converts these prediction errors into the residual utility measures defined in Section 3.2. Across both corpora, metadata labels provided measurable utility beyond text, but speaker identity provided larger utility and

left only small residual label utility after speaker control. The effect was clearest on L2-ARCTIC; VCTK showed the same ordering with smaller and less stable regional-accent utility. Thus, a metadata label can produce detectable average gains while offering little speaker-controlled utility, indicating that much of its apparent utility overlaps with speaker-associated structure. We next tested whether this ordering is preserved under likelihood-based scoring before turning to the diagnostics.

Likelihood-based residual utility. As a distributional robustness check, we additionally evaluated matched Gaussian predictors under held-out likelihood. Table 3 reports target-wise held-out Gaussian cross-entropy reductions from these predictors. The likelihood-based diagnostic preserved the same ordering. Speaker identity yielded the largest gain for nearly all target groups, metadata labels provided smaller but mostly positive gains beyond text, and metadata gains after speaker control were much smaller and occasionally negative. Again, the label effect was clearest on L2-ARCTIC, where L1 labels improved conditional likelihood beyond text for all target groups. VCTK regional-accent labels showed the same ordering with smaller gains. Thus, the speaker-dominant attribution was preserved under both point-prediction and probabilistic scoring.

4.3 Speaker-Held-Out Transfer

We separately evaluated speaker-held-out prediction to test whether metadata-associated regularities transfer to speakers unseen during training and validation. Because trained speaker embeddings are unavailable for unseen speakers without a separate enrollment mechanism, we compare only text-only and text+label predictors.

Metadata-conditioned transfer was limited and target-dependent. Label conditioning improved prominence in both corpora, with a larger gain on L2-ARCTIC, while effects on the remaining targets were small or mixed. In particular, L2-ARCTIC showed degradation on several non-prominence targets. Thus, metadata gains observed in the observed-speaker setting did not consistently transfer to unseen speakers under the current predictor. Full results are reported in Appendix B.6.

4.4 Label Specificity, Dependence, and Residual Structure

We next examine whether the observed metadata gains are specific to the supplied labels, whether

Model	Prom.	Boundary	PitchShape	RelDur
VCTK				
Text-only	0.327±0.001	0.301±0.002	1.445±0.009	0.180±0.001
Text+Rand.	0.332±0.002	0.314±0.002	1.450±0.009	0.180±0.001
Text+Acc.	0.322±0.001	0.298±0.001	1.433±0.003	0.179±0.001
L2-ARCTIC				
Text-only	0.413±0.001	0.373±0.001	1.556±0.002	0.228±0.001
Text+Rand.	0.417±0.001	0.380±0.001	1.586±0.002	0.230±0.001
Text+L1	0.392±0.001	0.363±0.001	1.499±0.002	0.222±0.001

Table 4: Random speaker-group control. Text+Rand. uses same-cardinality random speaker-level group labels. Values are mean MAE ± standard deviation; lower is better. Text+Rand. is averaged over five training seeds and three random group assignments per seed.

Test	Model	Prom.	Boundary	PitchShape	RelDur
VCTK					
True	Text+Acc.	0.322	0.298	1.433	0.179
Shuf.	Text+Acc.	0.335	0.303	1.474	0.184
True	Full	0.303	0.292	1.377	0.175
Shuf.	Full	0.319	0.298	1.426	0.179
L2-ARCTIC					
True	Text+L1	0.392	0.363	1.499	0.222
Shuf.	Text+L1	0.439	0.389	1.638	0.236
True	Full	0.376	0.351	1.410	0.218
Shuf.	Full	0.408	0.372	1.527	0.227

Table 5: Shuffled-label evaluation. Model parameters are fixed and only the supplied test-time metadata labels are permuted. Values are MAE averaged over five fixed random seeds; lower is better.

trained predictors actually depend on those labels, and how label-associated residual structure compares with speaker-associated structure. Because the metadata labels are assigned at the speaker level, a label-conditioned gain may reflect label-aligned prosodic structure, arbitrary coarse speaker grouping, or speaker-associated variation. We therefore combine complementary diagnostics. Random speaker-group controls test whether observed gains exceed arbitrary same-cardinality speaker partitions. Shuffled-label evaluation and label-swap sensitivity test whether fixed trained predictors change when the supplied metadata label is perturbed at test time. Residual-variance analysis and prompt-matched residual proximity then provide model-independent views of label- and speaker-associated residual structure.

Random speaker-group control. Table 4 reports the random speaker-group control. Text+Rand. uses random speaker-level groups with the same number of categories as the observed metadata la-

Model	Prom.	Boundary	PitchShape	RelDur
VCTK				
Text+Acc.	0.104	0.051	0.451	0.048
Full	0.106	0.053	0.425	0.040
L2-ARCTIC				
Text+L1	0.210	0.150	0.552	0.078
Full	0.172	0.128	0.493	0.063

Table 6: Label-swap sensitivity. Values are mean absolute prediction changes when the supplied metadata label is replaced by an alternative label while text is fixed; for the full model, speaker identity is also fixed. Larger values indicate greater dependence of the predictions on the supplied label.

Factor	Prom.	Boundary	PitchShape	RelDur
VCTK				
Label	0.011	0.018	0.012	0.002
Speaker	0.120	0.110	0.116	0.019
L2-ARCTIC				
Label	0.123	0.109	0.034	0.003
Speaker	0.184	0.206	0.138	0.013

Table 7: Residual variance explained over the full aligned subsets after removing prompt-level target means. Values are one-way η^2 for speaker identity and the metadata label: regional accent for VCTK and L1 for L2-ARCTIC.

bels, controlling for an added conditioning variable and arbitrary coarse speaker partitioning. On VCTK, Text+Rand. was worse than Text+Acc. for all targets and did not improve over text-only. On L2-ARCTIC, Text+Rand. was worse than both text-only and Text+L1 across all targets. Thus, same-cardinality random groups failed to reproduce the observed metadata gains. This shows that the gain is not reproduced by this arbitrary categorical grouping, but does not by itself establish a causal or purely linguistic accent- or L1-specific effect.

Shuffled-label evaluation. Table 5 reports shuffled-label evaluation, where model parameters are fixed and only the supplied test-time label is permuted. Across both corpora, shuffled labels increased MAE for both the text+label and full models. The degradation was largest for pitch shape and was especially pronounced on L2-ARCTIC, but it was visible across targets. This showed that trained predictors depended on the supplied label values at evaluation time rather than ignoring the label input.

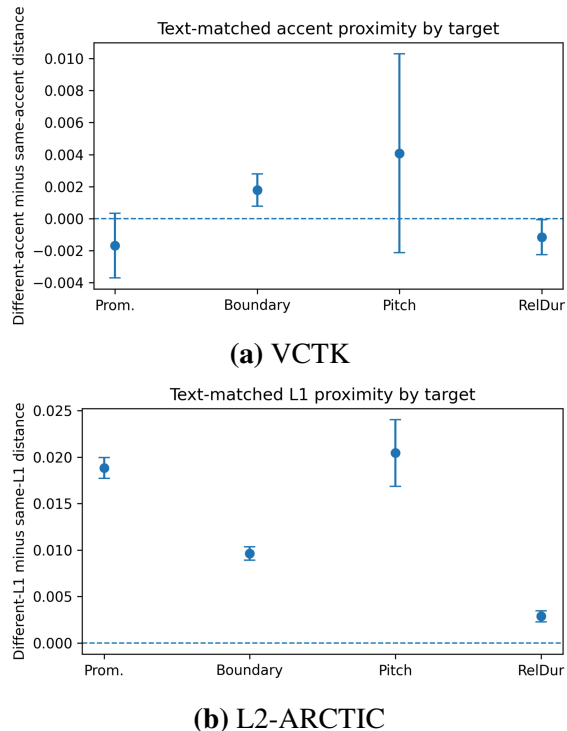


Figure 2: Prompt-matched label proximity in residual prosody. Bars show different-label minus same-label residual-prosody distance for utterances sharing the same prompt. Positive values indicate that same-label realizations are closer. The effect is clearer for L2-ARCTIC and weaker for VCTK, broadly consistent with the prediction-based decomposition.

Label-swap sensitivity. Table 6 reports label-swap sensitivity. We replaced the supplied metadata label while keeping the text fixed; in the full model, speaker identity was fixed as well. Across both corpora, swaps changed predictions for all targets, with the largest changes for pitch shape. The effect was larger on L2-ARCTIC, consistent with the stronger L1-conditioned gains above. VCTK also remained sensitive to regional-accent swaps in the full model, indicating that small speaker-controlled utility reflected limited held-out accuracy gain rather than complete label insensitivity.

Together, these perturbation results separate label dependence from label utility. Speaker-level metadata labels affected model predictions, but their marginal contribution to held-out accuracy was limited once speaker identity was available as a control. Because swaps create out-of-distribution speaker-label combinations, we interpret them as sensitivity diagnostics rather than counterfactual accent or L1 interventions. We next turn to model-independent diagnostics of residual prosodic structure.

Cond.	F_0	F_0	Energy	Duration
	RMSE↓	Corr.↑	Corr.↑	MAE↓
No pros.	1.647	0.890	0.799	0.531
Text	1.480	0.920	0.807	0.482
+Acc.	1.466	0.925	0.808	0.478
+Spk.	1.459	0.920	0.809	0.482
Full	1.461	0.924	0.809	0.467

Table 8: Auxiliary acoustic realization check on VCTK. No pros. denotes the fixed StyleTTS2 baseline without injected prosody plans; other conditions inject predicted prosody plans into the same backbone. Utterance-level standard deviations and additional metrics are reported in Appendix C.

Residual variance after prompt control. Table 7 compares label- and speaker-associated residual structure after subtracting prompt-level target means. Across both corpora, speaker identity explained more residual variance than the available speaker-level label for nearly all target groups. Because shared prompt-level structure had been removed, we use this diagnostic to compare the relative ordering of speaker- and label-associated residual structure rather than to estimate total prosodic variance explained. This model-independent diagnostic agrees with the prediction results: label-aligned residual structure was present, but speaker-associated structure was larger.

Prompt-matched residual proximity. Figure 2 gives a model-independent view of label-aligned residual structure: among utterances sharing a prompt, positive gaps mean that same-label realizations are closer in residual prosody space. The effect was clearest on L2-ARCTIC, while VCTK showed a weaker and target-dependent pattern.

4.5 Auxiliary Check: Acoustic Realization of Predicted Plans

Finally, we used a VCTK-only auxiliary check to test whether predicted plans induce measurable acoustic changes when injected into a fixed StyleTTS2 backbone.

Table 8 shows that predicted-plan injection improved F_0 error, F_0 /energy correlations, and duration error over the no-prosody baseline. This indicates that the predicted plans induced realizable acoustic variation rather than serving only as abstract regression targets. Differences among text-, metadata-, speaker-, and full-plan conditions were comparatively small, so we used this check only to validate acoustic realization of the target repre-

sentation, not to rank conditioning methods as TTS systems.

5 Conclusion

We presented a speaker-controlled attribution audit of speaker-level metadata utility in text-residual prosody prediction. Across VCTK and L2-ARCTIC, regional-accent and L1 labels provided measurable predictive utility beyond text, while speaker conditioning yielded larger gains and left only small additional label utility after speaker control. Likelihood-based scoring and complementary diagnostics supported the same attribution pattern. Speaker-held-out evaluation further showed that these metadata gains did not consistently transfer to unseen speakers across targets.

These findings motivate caution when interpreting predictive gains from metadata assigned at the speaker level. Such gains should not by themselves be treated as evidence of category-level prosodic effects; instead, they should be evaluated against broader speaker-associated structure and, where relevant, speaker-held-out transfer. The same audit principle may apply to other group labels attached to identifiable units.

Limitations

This work estimates predictive utility rather than causal accent or L1 effects. Speaker identity is used as a conservative control for broad speaker-associated structure and may absorb habitual prosody, pronunciation, speaking style, recording conditions, and other corpus-linked factors. Accordingly, small residual label utility after speaker control does not imply the absence of category-level effects.

Our prosody targets are automatically extracted acoustic measurements rather than perceptual ground truth, and speaker- or group-dependent extraction bias may remain. The four targets also cover only selected dimensions of prosody. Speaker-held-out evaluation tests transfer to unseen speakers within observed label categories, but not to unseen regional accents or L1s, and the observed transfer was target-dependent. Random-group and label-perturbation controls likewise test specific alternative explanations rather than ruling out all corpus confounds.

Future work should evaluate larger and more balanced corpora with multiple speakers per category, shared prompts, and better-controlled record-

ing conditions, together with additional prosodic representations and perceptual evaluation. Accent and L1 labels should not be interpreted as inherent properties of individuals or groups; speaker-controlled auditing is intended to reduce unsupported category-level interpretations of corpus-specific predictive gains.

Acknowledgements

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00557944) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201373, Artificial Intelligence Graduate School Program (Hanyang University)).

AI Usage Disclosure

During manuscript preparation, ChatGPT (OpenAI) was used only for language polishing, including grammar correction and clarity refinement.

References

- Matthew Aylett and Alice Turk. 2006. [Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei](#). *The Journal of the Acoustical Society of America*, 119(5):3048–3058.
- Paul Boersma and David Weenink. 2024. [Praat: Doing phonetics by computer](#). Computer program.
- Mara Breen, Evelina Fedorenko, Michael Wagner, and Edward Gibson. 2010. [Acoustic correlates of information structure](#). *Language and Cognitive Processes*, 25(7–9):1044–1098.
- Cynthia G. Clopper and Rajka Smiljanic. 2011. [Effects of gender and regional dialect on prosodic patterns in American English](#). *Journal of Phonetics*, 39(2):237–245.
- Jennifer Cole. 2015. [Prosody in context: A review](#). *Language, Cognition and Neuroscience*, 30(1–2):1–31.
- Mansi Dhamne, Sneha Raman, and Preeti Rao. 2025. [Predicting prosodic boundaries for children’s texts](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31875–31885, Suzhou, China. Association for Computational Linguistics.
- Esther Grabe and Brechtje Post. 2002. [Intonational variation in the British Isles](#). In *Speech Prosody*, pages 343–346.
- Esther Grabe, Brechtje Post, Francis Nolan, and Kimberley Farrar. 2000. [Pitch accent realization in four varieties of British English](#). *Journal of Phonetics*, 28(2):161–185.
- Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. [Introducing Parselmouth: A Python interface to Praat](#). *Journal of Phonetics*, 71:1–15.
- Sun-Ah Jun, editor. 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, Oxford.
- D. Robert Ladd. 2008. *Intonational Phonology*, 2 edition. Cambridge University Press.
- Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. 2023a. [Phoneme-level BERT for enhanced prosody of text-to-speech with grapheme predictions](#). In *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023b. [StyleTTS 2: Towards Human-Level Text-to-Speech through style diffusion and adversarial training with large speech language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 19594–19621. Curran Associates, Inc.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Zofia Malisz, Erika Brandt, Bernd Möbius, Yoon Mi Oh, and Bistra Andreeva. 2018. [Dimensions of segmental variability: Interaction of prosody and surprisal in six languages](#). *Frontiers in Communication*, 3:25.
- Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. [Montreal Forced Aligner: Trainable text-speech alignment using Kaldi](#). In *Proceedings of Interspeech 2017*, pages 498–502.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. [PyTorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems*, pages 8024–8035.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. [FiLM: Visual reasoning with a general conditioning layer](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. [FastSpeech 2: Fast and high-quality end-to-end text to speech](#). In *International Conference on Learning Representations*.

Scott Seyfarth. 2014. [Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation](#). *Cognition*, 133(1):140–155.

Rachel Smith and Tamara Rathcke. 2020. [Dialectal phonology constrains the phonetics of prominence](#). *Journal of Phonetics*, 78:100934.

Antti Suni, Sofoklis Kakouros, Martti Vainio, and Juraj Šimko. 2020. [Prosodic prominence and boundaries in sequence-to-sequence speech synthesis](#). In *Proceedings of the 10th International Conference on Speech Prosody*, pages 965–969.

Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. 2017. [Hierarchical representation and estimation of prosody using continuous wavelet transform](#). *Computer Speech & Language*, 45:123–136.

Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, and Martti Vainio. 2019. [Predicting prosodic prominence from text with pre-trained contextualized word representations](#). In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 281–290, Turku, Finland. Linköping University Electronic Press.

Christophe Veaux, Junichi Yamagishi, and Simon King. 2013. [The Voice Bank Corpus: Design, collection and data analysis of a large regional accent speech database](#). In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation*, pages 1–4.

Michael Wagner and Duane G. Watson. 2010. [Experimental and theoretical advances in prosody: A review](#). *Language and Cognitive Processes*, 25(7–9):905–945.

Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Gotlieb Wilcox, and Tamar I. Regev. 2023. [Quantifying the redundancy between prosody and text](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9765–9784, Singapore. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Trans-formers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. [L2-ARCTIC: A non-native English speech corpus](#). In *Proceedings of Interspeech 2018*, pages 2783–2787.

A Implementation, Data, and Analysis Details

A.1 Model Architecture and Conditioning

We provide implementation details for the RoBERTa-based prosody predictors, metadata embeddings, pooling operations, and prediction heads. Each predictor encodes the input text sequence with RoBERTa (Liu et al., 2019), and token representations are pooled to aligned word-level representations using token-to-word alignment maps. Syllable-level representations gather the corresponding word representation using syllable-to-word mappings, pool text-derived phone embeddings to syllables using phone-to-syllable maps, concatenate syllable-position embeddings, and append the broadcasted condition vector when conditioning is available. All conditioning settings receive the same text input and text-derived annotations; they differ only in whether the speaker-level metadata label and/or speaker identity are available. For VCTK, the label is regional accent; for L2-ARCTIC, the label is L1. The metadata label and speaker identity are injected through learned embeddings under the corresponding conditioning setting. All architectural choices are kept fixed across matched conditioning comparisons; only access to the metadata label and/or speaker identity changes.

A.2 Losses, Optimization, and Evaluation Metrics

Regression losses. For the main regression predictors, we trained word-level and syllable-level predictors with separate masked objectives:

$$\begin{aligned} L_{\text{word}} &= \lambda_{\text{prom}} L_{\text{prom}} + \lambda_{\text{bnd}} L_{\text{bnd}}, \\ L_{\text{syll}} &= \lambda_{\text{pitch}} L_{\text{pitch}} + \lambda_{\text{dur}} L_{\text{dur}}. \end{aligned} \quad (7)$$

All component losses in the main predictors are masked MSE losses computed only over valid aligned units. The word-level predictors were selected using validation L_{word} , and the syllable-level predictors were selected using validation L_{syll} . We used $\lambda_{\text{prom}} = \lambda_{\text{bnd}} = 1.0$, $\lambda_{\text{pitch}} = 0.5$, and $\lambda_{\text{dur}} = 5.0$ for all conditioning settings.

Target-wise likelihood diagnostic. For the likelihood-based diagnostic, we trained otherwise matched word-level and syllable-level Gaussian predictors with negative log-likelihood objectives. The conditioning settings, data splits, target masks, and target-specific loss weights were kept matched to the corresponding regression predictors. The NLL variant keeps the same mean prediction heads as the regression model and adds parallel heads that predict diagonal log-variances for each target dimension. For NLL models, L_{prom} , L_{bnd} , L_{pitch} , and L_{dur} denote masked diagonal-Gaussian negative log-likelihoods for the corresponding prominence, boundary-strength, pitch-shape, and relative-duration targets, computed only over valid aligned units. The predicted log-variances are clamped to $[-8, 5]$, and the constant Gaussian term is omitted because it is identical across matched models.

For the target-wise likelihood utilities reported in Table 3, we compute held-out Gaussian NLL separately for each target group, following the definitions in Section 3.2. That is, prominence, boundary strength, pitch shape, and relative duration are scored over their own valid target units before computing the corresponding within-seed NLL reductions.

Aggregate likelihood diagnostic. For the supplementary aggregate seed-level diagnostic, we additionally combine the selected word-level and syllable-level likelihood losses as

$$L_{\text{NLL}} = \lambda_{\text{prom}} L_{\text{prom}} + \lambda_{\text{bnd}} L_{\text{bnd}} + \lambda_{\text{pitch}} L_{\text{pitch}} + \lambda_{\text{dur}} L_{\text{dur}}. \quad (8)$$

Here, L_{prom} , L_{bnd} , L_{pitch} , and L_{dur} are averaged over their corresponding valid target units before aggregation; L_{pitch} averages the three pitch-shape dimensions. We use $\lambda_{\text{prom}} = \lambda_{\text{bnd}} = 1.0$, $\lambda_{\text{pitch}} = 0.5$, and $\lambda_{\text{dur}} = 5.0$, matching the relative target weights used for model selection and evaluation.

Both target-wise and aggregate likelihood-based utilities are computed as within-seed NLL reductions after re-evaluating validation-selected checkpoints on the fixed test split, using the same residual-utility logic as in Section 3.2. These likelihood-based results test whether the conditioning order observed with MAE is preserved under probabilistic scoring.

Optimization and checkpoint selection. Models were trained for 10 epochs, with checkpoints

selected on the validation split and evaluated once on the fixed test split. Unless otherwise stated, models were trained with batch size 8, AdamW optimization, learning rate 2×10^{-5} , weight decay 0.01, linear warmup ratio 0.06, gradient clipping at 1.0, and dropout 0.1.

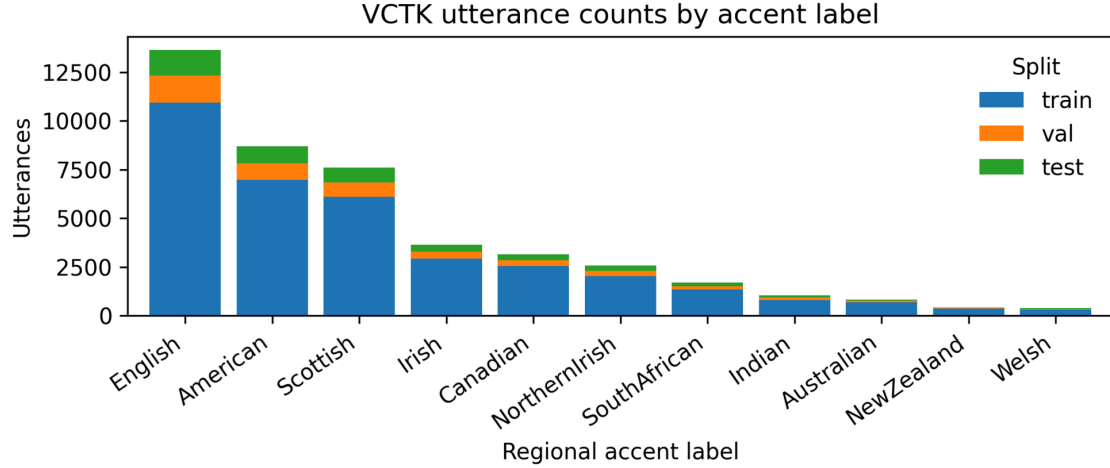
Evaluation metrics. We compute masked MAE, MSE, and Pearson correlation for word- and syllable-level targets. All metrics are computed only over valid aligned units. Main MAE-based prediction and residual-utility tables report averages over five fixed random seeds. For residual utility measures, positive values indicate error reduction relative to the corresponding baseline. Because the goal is matched comparison rather than absolute benchmark estimation, we use the same fixed test split for all ablations and interpret results as within-split estimates of residual predictive utility. The VCTK-only acoustic realization check uses a single synthesis run, with utterance-level variation reported in Appendix C.

Implementation environment. All models were implemented in PyTorch (Paszke et al., 2019) using Hugging Face Transformers (Wolf et al., 2020). Experiments were run on a single NVIDIA RTX 5090 GPU. Each model was trained for 10 epochs, and a typical run required approximately 80 minutes. Across the main regression experiments, this corresponds to approximately 53 GPU-hours for two corpora, four conditioning settings, and five random seeds. The predictor contains approximately 127M trainable parameters, including the pretrained RoBERTa encoder.

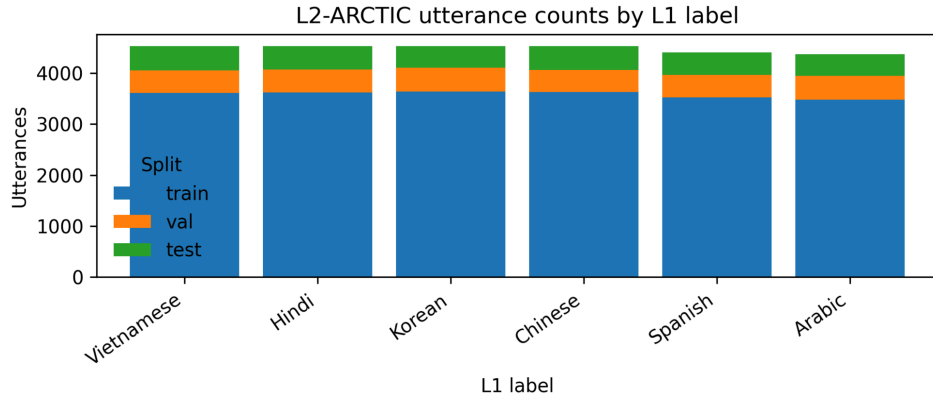
A.3 Datasets, Speaker-Level Labels, and Splits

Corpora and labels. We evaluated on aligned subsets of VCTK (Veaux et al., 2013) and L2-ARCTIC (Zhao et al., 2018) with available word- and syllable-level prosody targets. VCTK provides speaker-level regional-accent labels, whereas L2-ARCTIC provides speaker-level L1 labels. We treat these as corpus-specific speaker-level metadata labels rather than as instances of a shared linguistic variable. The predictor uses aligned word sequences, syllable/phone annotations, and speaker/label metadata.

Dataset licenses. We used both corpora under their respective research-use licenses. L2-ARCTIC is released under the Creative Commons



(a) VCTK regional-accent labels.



(b) L2-ARCTIC L1 labels.

Figure 3: Speaker-level label distributions under the main splits. Bars show utterance counts by speaker-level label, stacked by train, validation, and test split. VCTK regional-accent categories are imbalanced, while L2-ARCTIC L1 categories are balanced by speaker count.

Attribution–NonCommercial 4.0 license (CC BY-NC 4.0). The CSTR VCTK Corpus is released under the Open Data Commons Attribution License (ODC-By) v1.0. We use these corpora only for non-commercial research and cite the corresponding corpus papers.

Main utterance-level splits. The main experiments used fixed utterance-level train, validation, and test splits shared across all conditioning settings and seeds. Figure 3 shows the speaker-level label distributions under these main splits. For VCTK, the aligned subset contained 43,647 utterances from 107 speakers, split into 34,917 train, 4,364 validation, and 4,366 test utterances. For L2-ARCTIC, the aligned subset contained 26,866 utterances from 24 speakers, split into 21,493 train, 2,687 validation, and 2,686 test utterances. Val-

idation sets were used only for checkpoint selection, and test sets were evaluated only after checkpoint selection. Because speakers and prompts may recur across utterance-level splits, the main results estimate observed-speaker residual utility rather than speaker-independent generalization. All target-specific metrics were computed over valid aligned units after alignment and target masking.

Speaker-held-out diagnostic splits. For the speaker-held-out diagnostic, we constructed separate splits with disjoint training, validation, and evaluation speakers. For VCTK, 85 speakers provided 34,617 training utterances, 10 speakers provided 4,192 validation utterances, and 12 speakers provided 4,838 evaluation utterances. For L2-ARCTIC, 12 speakers provided 13,293 training utterances, 6 speakers provided 6,789 validation

Split	Utts.	Spk.	Lbl.
VCTK			
Main train	34,917	107	11
Main val.	4,364	107	11
Main test	4,366	107	11
Held-out train	34,617	85	11
Held-out val.	4,192	10	5
Held-out eval.	4,838	12	6
L2-ARCTIC			
Main train	21,493	24	6
Main val.	2,687	24	6
Main test	2,686	24	6
Held-out train	13,293	12	6
Held-out val.	6,789	6	6
Held-out eval.	6,784	6	6

Table 9: Dataset split statistics under the main and speaker-held-out splits. Labels denote regional-accent categories for VCTK and L1 categories for L2-ARCTIC. Validation splits are used only for checkpoint selection; main test and held-out eval. splits are reserved for final evaluation. The main split is an utterance-level split, so the same speaker pool appears across train, validation, and test. In the held-out diagnostic, evaluation speakers are unseen during both model fitting and checkpoint selection.

utterances, and 6 speakers provided 6,784 evaluation utterances. Evaluation speakers belonged to label categories observed among the training-side speakers. Thus, this diagnostic tests transfer to unseen speakers within observed label categories, not transfer to unseen regional accents or unseen L1s. Table 9 summarizes the main and speaker-held-out split statistics.

A.4 Prosody Target Extraction

We derive hierarchical prosody targets from forced-aligned speech. Word and syllable intervals are obtained from TextGrid alignments produced by Montreal Forced Aligner (McAuliffe et al., 2017). We remove non-speech intervals, including silence, pauses, punctuation-only labels, and empty labels.

Prominence and boundary strength. Word-level prominence and boundary strength are extracted from aligned speech using wavelet-based prosody representations (Sun et al., 2017). For each aligned word w_j , the wavelet extractor provides a raw prominence score m_j and a raw boundary score b_j from acoustic cues such as F0, energy, duration, and pausing. We use robust global-

normalized prominence and boundary targets:

$$\begin{aligned}\bar{m}_j &= \frac{m_j - \text{median}_{k \in \mathcal{W}_{\text{norm}}} m_k}{\text{IQR}_{k \in \mathcal{W}_{\text{norm}}} m_k}, \\ \bar{b}_j &= \frac{b_j - \text{median}_{k \in \mathcal{W}_{\text{norm}}} b_k}{\text{IQR}_{k \in \mathcal{W}_{\text{norm}}} b_k}.\end{aligned}\quad (9)$$

Here, $\mathcal{W}_{\text{norm}}$ denotes the word tokens used to estimate the robust normalization constants in the target files. If the interquartile range is numerically zero, we replace the scale with one. The resulting normalized values correspond to the prominence and boundary regression targets used in the main experiments.

Pitch shape. Frame-level F0 is extracted using Praat/Parselmouth with a 10 ms frame step and a 75–500 Hz pitch range (Boersma and Weenink, 2024; Jadoul et al., 2018). Unvoiced frames are excluded. Voiced F0 values are converted to semitones,

$$z_a = 12 \log_2 \frac{F0_a}{F0_{\text{ref}}} \quad (10)$$

where a indexes an F0 frame. The reference only affects the absolute offset because we subsequently subtract an utterance-level median. To reduce dependence on speaker-specific pitch height, we normalize voiced semitone values as

$$\tilde{z}_a = z_a - \text{median}_{\tau \in \mathcal{V}_u} z_\tau \quad (11)$$

where $\mathcal{V}_u$ is the set of voiced F0 frames in utterance u . For each syllable, we collect voiced frames within its interval. To reduce onset/offset and alignment noise, we crop the syllable by

$$m_i = \min(0.02, 0.15d_i) \quad (12)$$

seconds on both sides before collecting frames, falling back to the full interval if the cropped region is empty. We remove octave-jump-like frames by retaining only values satisfying

$$|\tilde{z}_a| \leq 12. \quad (13)$$

Let $\mathbf{v}_i = (\tilde{z}_{i,1}, \dots, \tilde{z}_{i,N_i})$ be the remaining utterance-normalized semitone contour for syllable σ_i . We linearly resample this contour to three points on normalized syllable time:

$$\mathbf{p}_i = \text{Interp}(\mathbf{v}_i, 3) = [p_{i,1}, p_{i,2}, p_{i,3}]. \quad (14)$$

The three values correspond to early, middle, and late relative F0 levels within the syllable. If only one voiced frame remains, the value is repeated; if

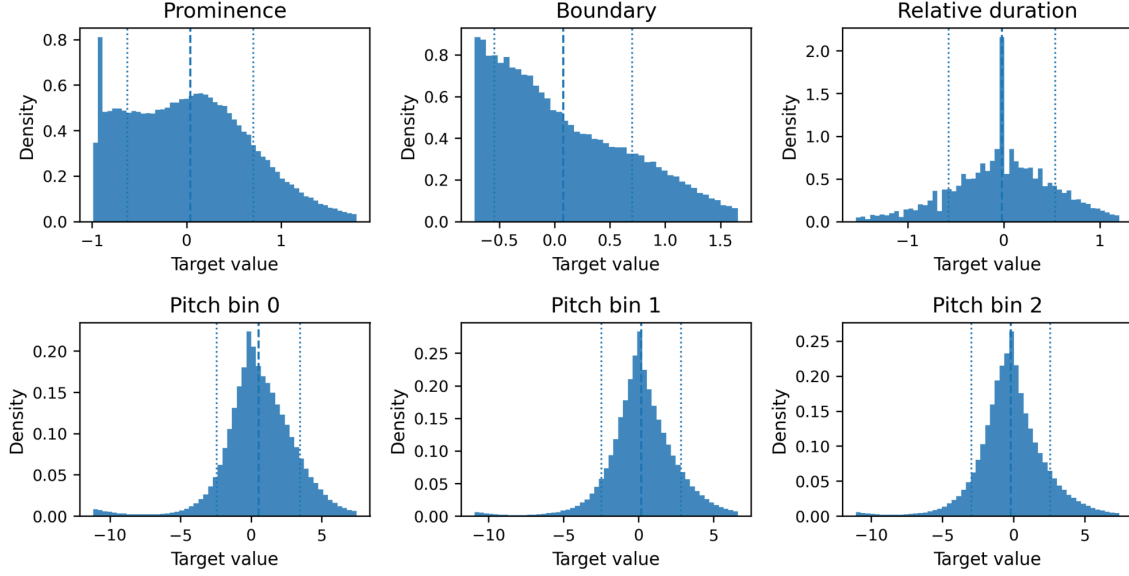


Figure 4: Distribution of extracted prosody targets after alignment and filtering. Dashed lines indicate means and dotted lines indicate one standard deviation.

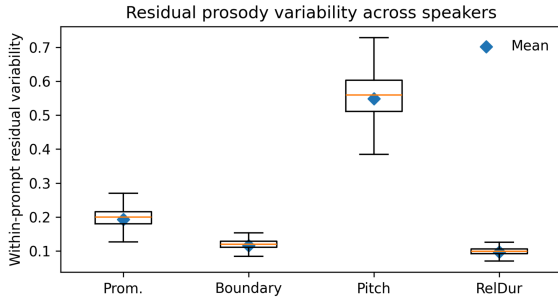


Figure 5: VCTK within-prompt residual prosody variability. For utterances sharing the same prompt, residual variability is computed from distances to prompt-level residual centroids. Pitch shape shows the largest within-prompt residual variability. The corresponding label-proximity panel is shown in Figure 2.

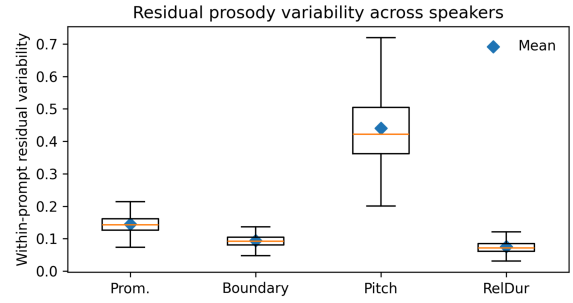


Figure 6: L2-ARCTIC within-prompt residual prosody variability. For utterances sharing the same prompt, residual variability is computed from distances to prompt-level residual centroids. Pitch shape shows the largest within-prompt residual variability. The corresponding label-proximity panel is shown in Figure 2.

no voiced frame remains, the pitch target is marked invalid and masked. We also mark pitch targets as reliable only when at least three voiced frames remain after filtering.

Relative duration. Let an utterance u contain valid syllables $\sigma_{1:S}$, where syllable σ_i has start time b_i , end time e_i , and duration $d_i = e_i - b_i$. We normalize syllable duration within each utterance to reduce sensitivity to global speaking rate. The relative-duration target is

$$r_i = \log \frac{d_i}{\text{median}_{j \in \mathcal{S}_u} d_j} \quad (15)$$

where $\mathcal{S}_u$ is the set of valid speech syllables in utterance u . This target captures local lengthening

or shortening relative to the utterance-level timing baseline.

Alignment and masking. For syllable–word alignment, syllables are associated with words using the syllable midpoint when it falls inside a word interval; otherwise, the word with maximum temporal overlap is used. All training losses and evaluation metrics are computed only over valid aligned units, i.e., word or syllable positions that survive alignment filtering and have both a text-derived representation and a corresponding non-missing prosody target. Units are masked when alignment information is missing, when duration is invalid, or, for pitch-shape targets, when insufficient voiced

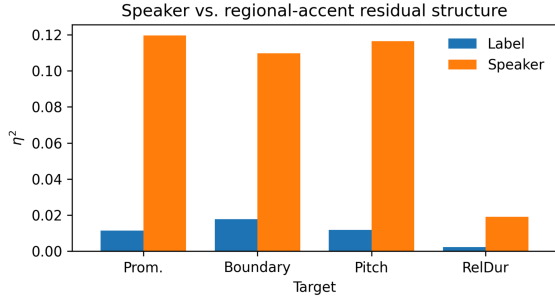


Figure 7: Speaker and regional-accent structure in VCTK text-residual prosody. Bars show variance explained (η^2) by speaker identity and regional-accent labels after removing prompt-level target means. Speaker-associated structure is larger than regional-accent structure across target groups, with relative duration showing very small explained variance overall.

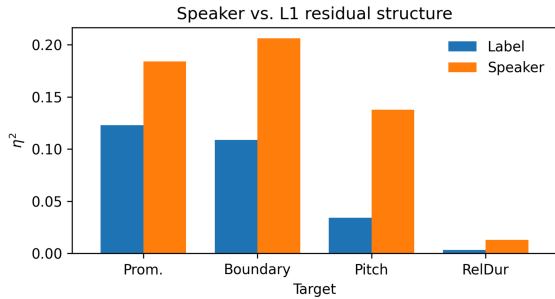


Figure 8: Speaker and L1 structure in L2-ARCTIC text-residual prosody. Bars show variance explained (η^2) by speaker identity and L1 labels after removing prompt-level target means. Speaker-associated structure is larger than L1-label structure for nearly all target groups, while both factors are near zero for relative duration.

frames are available. Figure 4 summarizes the resulting target distributions.

A.5 Descriptive Residual Structure

We include descriptive residual analyses that motivate the predictive utility experiments. These analyses are model-independent: they examine how residual prosody is structured after controlling for shared prompt-level content.

Figure 2 in the main text shows the label-proximity panels. Figures 5 and 6 additionally report within-prompt residual variability around prompt-level centroids. Across both corpora, pitch shape showed the largest within-prompt residual variability, supporting the view that text alone underdetermines prosodic realization. Figures 7 and 8 show residual-variance analyses after removing prompt-level target means. For utterances sharing the same prompt, we subtracted the prompt-level

Corpus	Prom.	Boundary	PitchShape	RelDur
VCTK	-0.0008	0.0010	0.0076	-0.0015
L2-ARCTIC	0.0185	0.0101	0.0557	0.0037

Table 10: Exact prompt-matched label proximity in residual prosody. For each prompt, we subtract the prompt-level target mean and compare pairwise residual-prosody distances only among utterances sharing that prompt. Values are different-label minus same-label distances, averaged over eligible prompts. Positive values indicate that same-label realizations are closer. The number of eligible prompts is 495 for VCTK and 1,132 for L2-ARCTIC.

target mean and computed one-way η^2 for speaker identity and the corresponding speaker-level label. Across both corpora, speaker identity explained more residual variance than the speaker-level label for nearly all target groups, with both factors near zero for relative duration. This descriptive result is consistent with the matched prediction results: after prompt-level structure was removed, text-residual prosody was more strongly organized by speaker identity than by the available speaker-level labels.

Exact prompt-matched label proximity. We additionally computed an exact prompt-matched proximity diagnostic, restricting pairwise comparisons to utterances that share the same prompt. For each prompt, we first subtracted the prompt-level target mean and then compared residual-prosody distances between same-label and different-label utterance pairs. Table 10 reports different-label minus same-label distance; positive values indicate that same-label realizations are closer in residual prosody space. The pattern is consistent with the main prediction results: L2-ARCTIC shows positive label proximity for all target groups, while VCTK is weaker and target-dependent.

B Supplementary Diagnostics and Robustness Results

This section reports additional diagnostics supporting the audit: uncertainty estimates, label specificity, label dependence, few-speaker robustness, and unseen-speaker transfer.

B.1 Residual Utility Uncertainty

We report paired utterance-level bootstrap confidence intervals and seed-level variation for residual utility estimates. Figure 9 visualizes the seed-level residual utilities across corpora and target groups. Tables 13 and 14 report VCTK and L2-ARCTIC

Seed	+Meta	+Spk.	Full	Meta Spk.
VCTK				
7	0.077	0.297	0.312	0.015
13	0.061	0.261	0.313	0.052
29	0.086	0.273	0.305	0.032
37	0.085	0.282	0.293	0.011
73	0.022	0.205	0.224	0.019
Mean	0.066	0.264	0.290	0.026
L2-ARCTIC				
7	0.167	0.341	0.364	0.023
13	0.194	0.357	0.381	0.024
29	0.193	0.347	0.377	0.030
37	0.192	0.384	0.404	0.020
73	0.197	0.362	0.385	0.023
Mean	0.189	0.358	0.382	0.024

Table 11: Seed-level aggregate likelihood-based metadata utility diagnostic. Values are within-seed reductions in aggregate held-out NLL; positive values indicate better likelihood. +Meta, +Spk., Full, and Meta|Spk. denote metadata-label conditioning, speaker conditioning, both, and metadata gain after speaker control, respectively.

Accent label	ID	Spk.	Test utts.
Welsh	10	1	46
NewZealand	6	1	45
Australian	1	2	71
Indian	4	3	116
SouthAfrican	9	4	174
NorthernIrish	7	6	279
Canadian	2	8	310
Irish	5	9	337

Table 12: Few-speaker VCTK accent groups. We define few-speaker groups as regional accent labels with fewer than ten speakers in the full speaker set. The table reports speaker counts in the full speaker set and utterance counts in the main VCTK test split.

bootstrap intervals, respectively. Because the bootstrap operates over utterances, these intervals use utterance-averaged paired errors and may differ slightly from the unit-averaged seed-level utilities in Tables 15 and 16. Tables 15 and 16 report the corresponding seed-level mean, standard deviation, and positivity counts. These statistics are intended as robustness checks for the matched predictive comparisons.

B.2 Seed-Level Aggregate Likelihood-Based Diagnostic

Table 11 reports seed-level reductions in aggregate held-out NLL. This aggregate diagnostic is supplementary to the target-wise likelihood utilities reported in Table 3. Positive values indicate better

held-out likelihood relative to the corresponding baseline. As in the target-wise and MAE-based results, speaker conditioning gave larger gains than metadata conditioning, and the remaining metadata gain after speaker control was positive but small.

B.3 Random Speaker-Group Control

Tables 17 and 18 report seed variance for the random speaker-group control in Table 4. Text-only and Text+label correspond to the five-seed main evaluations, where the label is regional accent for VCTK and L1 for L2-ARCTIC. Text+Rand. averages three random speaker-group assignments within each of the same five training seeds.

B.4 Perturbation Diagnostic Seed Variation

Tables 19 and 20 report seed-level variation for the shuffled-label and label-swap diagnostics in Tables 5 and 6. The mean values match the main tables; the appendix adds standard deviations over the same five fixed random seeds. These diagnostics are used to test whether trained predictors depend on the supplied metadata label input, rather than to estimate additional residual utility.

B.5 Few-Speaker VCTK Accent Subset

We additionally evaluate the main VCTK conditioning order on few-speaker accent groups, defined as regional-accent labels with fewer than ten speakers in the full speaker set. Table 12 lists the resulting groups, and Table 21 shows that the main ordering is preserved on this subset: accent conditioning improves over text-only, but speaker conditioning yields larger gains. This subset is a robustness check, not a separate estimate of regional-accent effects. We do not construct an analogous L2-ARCTIC subset because its six L1 categories are balanced at four speakers per category.

B.6 Speaker-Held-Out Transfer Diagnostic

Table 22 reports the speaker-held-out transfer diagnostic with seed variation. Evaluation speakers are unseen during training and validation, so trained speaker embeddings cannot be used. We therefore compare only text-only and text+label predictors. The results were small and mixed across targets: label conditioning improved prominence in both corpora, had little or no effect on VCTK boundary and pitch shape, and worsened several L2-ARCTIC non-prominence targets. We therefore interpret this diagnostic as a transfer check rather than the primary utility estimate.

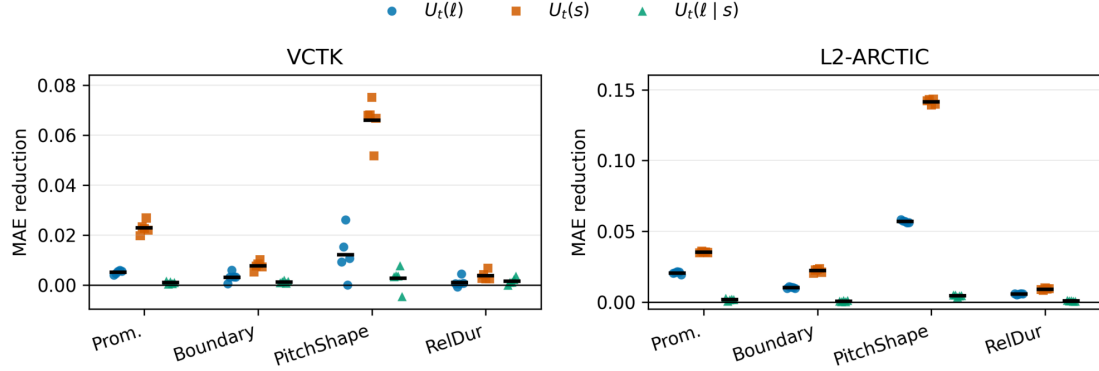


Figure 9: Seed-level residual utility estimates across corpora and target groups. Each point is one random seed, and horizontal black markers show the mean across five seeds. Utilities are MAE reductions relative to the corresponding matched baseline: $U_t(\ell)$ measures metadata-label utility beyond text, $U_t(s)$ measures speaker-identity utility beyond text, and $U_t(\ell | s)$ measures metadata-label utility after speaker control. The two panels use separate y-axis scales for readability. Although metadata labels provide positive residual utility beyond text for most targets, the seed-level pattern shows that speaker identity accounts for larger gains and that residual metadata utility after speaker control remains small.

Target	$U_t(\ell)$	$U_t(s)$	$U_t(\ell s)$
Prom.	0.0054 [0.0046, 0.0063]	0.0239 [0.0222, 0.0255]	0.0010 [0.0006, 0.0014]
Boundary	0.0033 [0.0027, 0.0039]	0.0080 [0.0069, 0.0090]	0.0013 [0.0009, 0.0016]
PitchShape	0.0126 [0.0103, 0.0148]	0.0679 [0.0626, 0.0732]	0.0021 [0.0008, 0.0034]
RelDur	0.0010 [0.0006, 0.0014]	0.0037 [0.0032, 0.0041]	0.0015 [0.0012, 0.0018]

Table 13: Paired utterance-level bootstrap 95% confidence intervals for VCTK residual utility estimates. Here, ℓ denotes speaker-level regional accent. For each bootstrap sample, held-out utterances are sampled with replacement and residual utilities are recomputed from paired utterance-averaged MAE differences across matched conditioning settings. Values report the five-seed mean utility with 95% confidence intervals from 10,000 bootstrap resamples. Positive values indicate lower held-out MAE.

C Auxiliary Synthesis Realization

We used a pretrained StyleTTS2 backbone (Li et al., 2023b) as an auxiliary realization interface on VCTK to test whether injected prosody targets produced measurable acoustic differences in synthesized speech. This evaluation checks whether the predicted target space is acoustically active when injected into a synthesis model, rather than serving as an accent-conversion or synthesis-quality benchmark. Wavelet-derived prominence and boundary features have also been used as explicit prosodic controls in speech synthesis (Sun et al., 2020).

Conditioning. Following FiLM-style modulation (Perez et al., 2018), each prosody target sequence is mapped to a conditioning vector q and used to modulate the PL-BERT (Li et al., 2023a) representation inside the fixed StyleTTS2 backbone (Li et al., 2023b): $\tilde{h}_t = (1 + \alpha\gamma(q)) \odot h_t + \alpha\beta(q)$, where h_t denotes the PL-BERT hidden representation at token position t , $\gamma(\cdot)$ and $\beta(\cdot)$ are learned transformations, and α controls modulation

strength. The modulation strength is fixed across realization conditions.

Conditions and metrics. All realization conditions use the same pretrained VCTK StyleTTS2 backbone and the same speaker/accent conditioning; they differ only in whether a prosody plan is injected and, when injected, which matched predictor provides that plan. We compare no-prosody synthesis with injection of predicted prosody targets from the text-only, text+accent, text+speaker, and full VCTK predictors. We report pitch, energy, and duration metrics as acoustic correlates of the predicted prosody targets.

Results. Table 23 reports the auxiliary realization results with utterance-level standard deviations. The same mean results are summarized in Table 8. The large utterance-level variance shows that these metrics should be interpreted as an auxiliary realization check rather than as a system-level TTS comparison.

Target	$U_t(\ell)$	$U_t(s)$	$U_t(\ell s)$
Prom.	0.0213 [0.0190, 0.0236]	0.0362 [0.0334, 0.0390]	0.0018 [0.0012, 0.0023]
Boundary	0.0102 [0.0086, 0.0119]	0.0214 [0.0193, 0.0236]	0.0006 [0.0002, 0.0011]
PitchShape	0.0565 [0.0501, 0.0627]	0.1406 [0.1323, 0.1492]	0.0040 [0.0030, 0.0050]
RelDur	0.0055 [0.0048, 0.0062]	0.0092 [0.0083, 0.0101]	0.0007 [0.0005, 0.0010]

Table 14: Paired utterance-level bootstrap 95% confidence intervals for L2-ARCTIC residual utility estimates. Here, ℓ denotes speaker-level L1 metadata. For each bootstrap sample, held-out utterances are sampled with replacement and residual utilities are recomputed from paired utterance-averaged MAE differences across matched conditioning settings. Values report the five-seed mean utility with 95% confidence intervals from 10,000 bootstrap resamples. Positive values indicate lower held-out MAE.

Target	$U_t(\ell)$	Pos.	$U_t(s)$	Pos.	$U_t(\ell s)$	Pos.
Prom.	0.0050±0.0008	5/5	0.0229±0.0026	5/5	0.0009±0.0005	5/5
Boundary	0.0031±0.0019	5/5	0.0077±0.0018	5/5	0.0012±0.0005	5/5
PitchShape	0.0122±0.0095	4/5	0.0659±0.0086	5/5	0.0027±0.0045	4/5
RelDur	0.0010±0.0020	4/5	0.0037±0.0019	5/5	0.0015±0.0014	4/5

Table 15: Seed-level uncertainty for VCTK residual utility estimates. Utility values are reported as mean $\pm$ standard deviation across five fixed random seeds. Here, ℓ denotes speaker-level regional accent. The Pos. columns report the number of seeds for which the corresponding utility is positive. These statistics are intended as robustness checks rather than formal significance tests.

Target	$U_t(\ell)$	Pos.	$U_t(s)$	Pos.	$U_t(\ell s)$	Pos.
Prom.	0.0205±0.0009	5/5	0.0353±0.0005	5/5	0.0017±0.0007	5/5
Boundary	0.0101±0.0005	5/5	0.0221±0.0014	5/5	0.0006±0.0004	5/5
PitchShape	0.0568±0.0009	5/5	0.1414±0.0018	5/5	0.0044±0.0007	5/5
RelDur	0.0055±0.0003	5/5	0.0091±0.0007	5/5	0.0008±0.0003	5/5

Table 16: Seed-level uncertainty for L2-ARCTIC residual utility estimates. Utility values are reported as mean $\pm$ standard deviation across five fixed random seeds. Here, ℓ denotes speaker-level L1 metadata, and s denotes speaker identity. The Pos. columns report the number of seeds for which the corresponding utility is positive.

Model	Prom.	Boundary	PitchShape	RelDur
VCTK				
Text-only	0.327±0.001	0.301±0.002	1.445±0.009	0.180±0.001
Text+Rand.	0.332±0.002	0.314±0.002	1.450±0.009	0.180±0.001
Text+Accent	0.322±0.001	0.298±0.001	1.433±0.003	0.179±0.001

Table 17: Seed-level variation for the VCTK random speaker-group control in Table 4. Values are mean MAE $\pm$ standard deviation. Text+Rand. averages three random group assignments within each of five training seeds.

Model	Prom.	Boundary	PitchShape	RelDur
L2-ARCTIC				
Text-only	0.413±0.001	0.373±0.001	1.556±0.002	0.228±0.001
Text+Rand.	0.417±0.001	0.380±0.001	1.586±0.002	0.230±0.001
Text+L1	0.392±0.001	0.363±0.001	1.499±0.002	0.222±0.001

Table 18: Seed-level variation for the L2-ARCTIC random speaker-group control in Table 4. Values are mean MAE $\pm$ standard deviation. Text+Rand. averages three random group assignments within each of five training seeds.

Test	Model	Prom.	Boundary	PitchShape	RelDur
VCTK					
True	Text+Acc.	0.322±0.001	0.298±0.001	1.433±0.003	0.179±0.001
Shuf.	Text+Acc.	0.335±0.001	0.303±0.001	1.474±0.005	0.184±0.001
True	Full	0.303±0.002	0.292±0.001	1.377±0.003	0.175±0.002
Shuf.	Full	0.319±0.002	0.298±0.001	1.426±0.006	0.179±0.001
L2-ARCTIC					
True	Text+L1	0.392±0.001	0.363±0.001	1.499±0.002	0.222±0.001
Shuf.	Text+L1	0.439±0.003	0.389±0.001	1.638±0.005	0.236±0.001
True	Full	0.376±0.001	0.351±0.001	1.410±0.001	0.218±0.001
Shuf.	Full	0.408±0.002	0.372±0.003	1.527±0.008	0.227±0.001

Table 19: Seed-level variation for the shuffled-label evaluation in Table 5. Model parameters are fixed and only the supplied test-time metadata labels are permuted. Values are mean MAE $\pm$ standard deviation over five fixed random seeds; lower is better.

Model	Prom.	Boundary	PitchShape	RelDur
VCTK				
Text+Acc.	0.104±0.009	0.051±0.006	0.451±0.018	0.048±0.001
Full	0.106±0.011	0.053±0.007	0.425±0.047	0.040±0.002
L2-ARCTIC				
Text+L1	0.210±0.012	0.150±0.007	0.552±0.015	0.078±0.002
Full	0.172±0.004	0.128±0.008	0.493±0.005	0.063±0.002

Table 20: Seed-level variation for the label-swap sensitivity diagnostic in Table 6. Values are mean absolute prediction changes $\pm$ standard deviation over five fixed random seeds when the supplied metadata label is replaced by an alternative label while text is fixed; for the full model, speaker identity is also fixed. Larger values indicate greater dependence of the predictions on the supplied label.

Model	Prom.	Boundary	PitchShape	RelDur
Text-only	0.338±0.001	0.304±0.001	1.530±0.009	0.184±0.001
+Acc.	0.331±0.001	0.300±0.001	1.510±0.001	0.182±0.001
+Spk.	0.316±0.002	0.297±0.001	1.439±0.005	0.180±0.001
Full	0.314±0.002	0.295±0.001	1.432±0.005	0.179±0.001

Table 21: Few-speaker VCTK accent subset results within the main test split. The subset includes regional-accent labels with fewer than ten speakers in the full speaker set, covering 1,378 test utterances from 34 speakers. Values are mean MAE $\pm$ standard deviation over five fixed seeds; lower is better.

Model	Prom.	Boundary	PitchShape	RelDur
VCTK				
Text-only	0.325±0.001	0.318±0.004	1.307±0.006	0.178±0.001
Text+Accent	0.323±0.001	0.317±0.004	1.307±0.004	0.177±0.001
L2-ARCTIC				
Text-only	0.421±0.0005	0.381±0.001	1.598±0.003	0.236±0.001
Text+L1	0.413±0.002	0.382±0.004	1.603±0.008	0.237±0.003

Table 22: Speaker-held-out transfer diagnostic. Evaluation speakers are unseen during training and validation, so only text-only and text+label predictors are compared. Values are mean MAE $\pm$ standard deviation over five fixed seeds; lower is better.

Condition	F_0 RMSE↓	F_0 Corr.↑	Energy RMSE↓	Energy Corr.↑	Dur. MAE↓
No prosody	1.647±1.165	0.890±0.238	1.176±0.244	0.799±0.078	0.531±0.509
Text-only predictor	1.480±0.972	0.920±0.141	1.148±0.200	0.807±0.072	0.482±0.483
Text+Acc. predictor	1.466±0.856	0.925±0.134	1.146±0.203	0.808±0.071	0.478±0.471
Text+Spk. predictor	1.459±0.779	0.920±0.140	1.143±0.197	0.809±0.069	0.482±0.483
Full predictor	1.461±0.790	0.924±0.132	1.138±0.198	0.809±0.070	0.467±0.472

Table 23: VCTK auxiliary synthesis realization check. All synthesis conditions are evaluated on the same evaluation samples. No prosody denotes the same StyleTTS2 backbone with speaker/accent conditioning but without injected prosody targets. The remaining conditions inject predicted prosody targets from the matched VCTK predictors. We report mean $\pm$ standard deviation over utterances. Lower F_0 RMSE, energy RMSE, and duration MAE are better; higher F_0 and energy correlations are better.