A Fixed-Text Diagnostic for Selector-Aligned Source-Acoustic Prompting in Speech-to-Speech Translation

Anonymous authors
Paper under double-blind review

Abstract

Speech-to-speech translation can preserve target words while changing delivery. We examine a source-acoustic policy that measures source speech to select a target-language prompt. Its selector-aligned score uses the same four acoustic summaries. We introduce a counterfactual protocol that fixes the text-to-speech (TTS) input, renderer, candidate bank, decoding, and utterance identifier while changing only prompt policy. We test 160 read-speech utterances from the Few-shot Learning Evaluation of Universal Representations of Speech (FLEURS) validation set across English-to-German, English-to-French, English-to-Dutch, and English-to-Turkish. We compare source-acoustic prompting with function-label-only prompting, one recorded shuffled-source allocation, and three recorded random allocations. On the selector-aligned score, source-acoustic prompting exceeds the recorded shuffled-source allocation but is inconclusive against the function-label and random controls. Against function labels, the score comparing generated speech with held-out target references has a lower point estimate in all four language pairs. Lower values mean less acoustic similarity. The English-to-Dutch interval remains inconclusive. The duration score is higher for English-to-German but lower in the other three pairs, so it does not show a stable timing benefit. A stricter automatic audit does not use the selector's four stored summary values when it scores the outputs. In that audit, only the pooled pitch contrast against shuffled source is positive, and we do not treat that isolated contrast as standalone evidence. Energy, duration, and pause remain inconclusive, and some Dutch contrasts still have no paired support. On the 80-item support-screened WavLM follow-up, frozen embeddings are lower against function labels and inconclusive against shuffled source; 29 of its IDs overlap the pilot, so it is not an independent-cohort replication. The predeclared positive rule requires both contrasts to be above zero and is not met. Together, these automatic checks show movement in the selector's preferred signal, not reliable improvement in held-out target speech.

1 Introduction

Speech-to-speech translation (S2ST) has to preserve more than lexical content. Pitch, energy, voicing, and pauses can change how an utterance is heard even when its translated text is unchanged. Recent work therefore studies expressive and interactional speech translation alongside lexical adequacy (Seamless Communication et al., 2023a;b; Jiang et al., 2026; Koudounas et al., 2026).[1]

Source-acoustic prompting is a tempting response: use measurements of source speech to select a target-language prompt, then render fixed translated text. This paper treats that idea as a stripped-down audit target, not as a comparison with the strongest source-conditioned method. The setup captures a basic heuristic suggested by prompt retrieval in expressive TTS and source-conditioned S2ST: match source and prompt acoustics, then ask whether

[1] An anonymous source-code archive accompanies the supplementary materials.

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.

any apparent gain survives controlled alternatives (Xue et al., 2024; Ulgen et al., 2026; Seamless Communication et al., 2023a). A higher acoustic score after this intervention can look like evidence of delivery transfer. That reading can fail when the score measures the same feature family that drove selection. In that case, the policy may move its own preferred signal while leaving open whether generated speech matches a held-out target reference or preserves what listeners hear.

We ask a narrow question: with the translated TTS input and the renderer fixed, what does a source-acoustic prompt policy change under controlled alternatives? We compare one source-acoustic policy with a weak function-label policy, a shuffled-source control, and three random prompt allocations. Every comparison keeps the utterance identifier, translated text, candidate bank, renderer checkpoint, decoding configuration, and generation seed fixed. Here fixed text refers to the renderer input string; generated waveforms are not verified to contain identical words across prompt policies. We then read the outputs through seven automatic score families. These are a selector-aligned source-summary score that compares generated speech with the source on the selector's four acoustic fields, a target-reference summary, a target-duration similarity, field-masked target-frame diagnostics, a stricter summary-exclusion audit, an appendix-only summary-invariant diagnostic, and frozen pretrained-representation similarity. None is a listener-level measure.

The result is not a consistent target-side gain. The selector exceeds the recorded shuffled-source allocation on its aligned score, but its contrasts with function-label-only and random allocations are inconclusive. Against function labels, the target-reference score is lower in all four language pairs, meaning the source-acoustic outputs are less acoustically similar to the held-out target references; English–Dutch remains inconclusive. The target-duration score is higher in English–German and lower in the other three pairs, so it offers no stable timing benefit. The stricter summary-exclusion audit only shows a positive shuffled-source contrast on pitch projection; its energy, duration, and pause contrasts remain inconclusive or negative, and some Dutch contrasts are unsupported because no shared identifiers retain finite projected scores in both compared conditions. The 80-item support-screened WavLM follow-up shares 29 utterance IDs with the pilot and is not an independent replication. With complete paired coverage, its function-label contrast is below zero and its shuffled-source contrast is inconclusive; the pooled negative contrast is concentrated in English–Turkish and is not evenly shared across pairs. This counterfactual pattern motivates the broader methodological lesson of the paper: selector-aligned improvement does not establish delivery preservation. The present study demonstrates that lesson on one heuristic over read-speech FLEURS items, not on the full family of source-conditioned methods.

Our contributions are:

- a fixed-text prompt-policy protocol that separates prompt effects from ASR, MT, renderer, and variation in the TTS input text;
- a four-language, utterance-paired pilot with source-acoustic, function-label-only, shuffled-source, and three random policies;
- a direct result in which selector-aligned movement is not accompanied by a consistent target-reference improvement in this setting;
- a protocolized 80-item WavLM follow-up that does not meet its positive automatic decision rule; 29 utterance IDs overlap with the pilot, so it is not an independent replication; and
- a stricter summary-exclusion reanalysis that does not read the selector's stored scalar summaries; its narrow controlled-perturbation check supports the scorer's intended response but leaves unsupported Dutch contrasts and no complete feature-disjoint target-side endpoint.

2 Why a fixed-text counterfactual is needed

Full S2ST evaluation mixes recognition, translation, generation, and prompt selection. A delivery-sensitive score can move because words, renderer, or prompt changed. Existing

Fixed: source utterance, translated text, prompt bank, renderer, decoding, and utterance identifier.

Intervention: select a prompt from source acoustics, weak function labels, a shuffled source profile, or a recorded random allocation.

Separate readouts: source-summary similarity shares the selector's feature family; target-reference similarity compares generated speech with held-out target reference speech; duration similarity reads target and generated durations only; stricter summary-exclusion diagnostics derive residual pitch, energy, and pause scores without reading the selector's four stored scalar fields.

Boundary: no automatic readout measures listener preference, naturalness, pragmatic-function preservation, or general delivery transfer.

Figure 1: Fixed-text counterfactual. Only the prompt policy changes. The aligned source-summary score tracks the selector's intended signal, while the remaining readouts test target-side behavior under the same fixed text and renderer.

benchmarks make delivery important, but full-system comparisons cannot isolate prompt-conditioned rendering after translation (Ward et al., 2026; An et al., 2026; Koudounas et al., 2026). Our study treats prompt selection as an intervention within a fixed cascade. It does not train a new model or rank S2ST architectures.

Prompt retrieval is already a practical conditioning mechanism in prompt-based TTS (Xue et al., 2024), while direct S2ST work studies prosody transfer through representations and generation mechanisms (Ulgen et al., 2026; Seamless Communication et al., 2023a). This paper proposes neither a new retrieval model nor a general prosody-transfer system. Its narrower contribution is a reusable counterfactual audit: holding text and rendering fixed exposes whether an apparent improvement survives controls that change only the prompt-policy input and whether the readout overlaps the selector. More broadly, the paper argues that prompt-selection results need paired controls and target-side checks that are not mechanically coupled to the selector. Figure 1 summarizes the fixed inputs, the sole prompt-policy intervention, the selector-aligned readout, and the automatic target-side diagnostics; it frames why an aligned score is not independent delivery evidence.

3 Protocol

For source speech x, fixed translated text y, prompt c, and generated target speech $\hat{z}_c$, the cascade is

$$y = \mathrm{MT}(\mathrm{ASR}(x)), \qquad \hat{z}_c = \mathrm{TTS}(y, c). \tag{1}$$

All policies share x, y, the candidate bank, the TTS checkpoint, decoding configuration, and generation seed. Only prompt selection changes. Here "fixed text" refers to the TTS input string y; it does not certify lexical identity of the generated speech across prompt policies.

Policies. Source-acoustic selection compares the item's source profile with candidate target-prompt profiles and never reads the item's target-reference acoustics. It uses summary-mode similarity over four scalar fields—mean pitch, RMS energy, voiced-frame ratio, and pause count—and score $s = a$, where a is the four-field acoustic-summary similarity below. Function-label-only selection uses score $s = \ell$, the Jaccard overlap between the query item's existing weak pragmatic_labels set and a candidate prompt's set, with acoustic weight zero; it is a text-derived proxy, not a validated pragmatic selector. Shuffled-source selection applies the source-acoustic policy after one deterministic within-language permutation of source profiles. Three recorded random policies assign one eligible prompt to each item using allocation seeds 20260918, 20260919, and 20260920 (A/B/C, respectively). In all conditions, the query item itself and candidates sharing its recorded document identifier are excluded; ties resolve deterministically by candidate identifier. Speaker-disjointness was not verified, and the current prompt-bank audit finds no populated speaker_id field, so overlap rates cannot be estimated from this record.

Readouts. For nonnegative scalar field q, its pairwise similarity is $1 - |q_1 - q_2| / \max(q_1, q_2, 10^{-12})$. The source and target-reference summaries average four field-wise values: mean pitch, root-mean-square energy, voiced-frame ratio, and pause count. The first three use the scalar rule; pause count uses $1 - |n_1 - n_2| / \max(1, n_1, n_2)$. The source summary compares source with generated speech; the target-reference summary substitutes the held-out target-language FLEURS waveform paired to the same utterance identifier. The selector uses the same source-summary family, so that endpoint is aligned with selection. Separately, target-duration similarity applies the scalar rule to the two waveform durations only. An implementation-level generated-WAV unit test holds duration fixed while changing pitch, amplitude, an internal silence span, and channel count; all four variants give duration similarity 1, while a waveform with half the duration gives 0.5. This confirms direct-field behavior, not endpoint validity. Duration can still covary with pausing in natural speech, so it is an automatic timing diagnostic, not a validated proxy for perceived intent, naturalness, preference, or delivery transfer.

The field-masked reanalysis compares centered, time-normalized target and generated log-pitch shapes, log-RMS shapes outside detected pauses, and pause-boundary F1. It does not read the selector's four stored summary fields: mean pitch, mean RMS, voiced-frame ratio, or pause count. The frame-level pitch and RMS values can nevertheless encode related information, so this is field-masked, not feature-disjoint. Synthetic profile tests move one designated pitch, energy, pause, or duration factor at a time and leave the other three scores fixed. A waveform without enough voiced or non-pause frames has a null shape score; each contrast then uses identifiers finite in both conditions for that metric and reports its own denominator. This is a measurement audit, not a complete delivery endpoint.

A stricter summary-exclusion reanalysis goes further. It does not read the selector's four stored scalar fields—mean pitch, mean RMS energy, voiced-frame ratio, or pause count—when constructing score inputs. It derives centered, resampled, DC-excluded residual vectors for log pitch, log RMS outside pauses, and pause occupancy, compares target and generated vectors with rescaled cosine similarity, and scores waveform duration separately. Synthetic profile tests again move the designated pitch, energy, pause, and duration factors while leaving the others fixed. A separate seven-case controlled validator reports that designated pitch, energy, duration, and pause perturbations change their intended scores, while stored-summary-only and nuisance-only changes leave the reported scores unchanged. This synthetic check supports the scorer's stated access pattern; it does not establish complete real-data coverage or an independent confirmation endpoint. Each metric still uses only identifiers finite in both compared conditions, and the audit marks a contrast unsupported when no shared identifiers retain finite scores in both compared conditions instead of imputing a value.

As a separate learned automatic audit, we freeze WavLM Base+ (resolved revision 4c66d4806a428f2e922ccfa1a962776e232d487b) and compare mean-pooled target-reference embedding similarity. This model is never exposed to prompt selection. It remains an automatic pretrained representation, so we interpret its scores from the direction and consistency of paired-control contrasts; no external perceptual scale is available, and this study has no repeated-synthesis or oracle calibration anchor for absolute magnitude.

Data and uncertainty. We use deterministic 40-item subsets of the FLEURS validation index (Conneau et al., 2022) for English–German, English–French, English–Dutch, and English–Turkish ($N = 160$). Every policy uses the same 40 utterance identifiers within each language pair. Each of the 24 language-pair-by-policy output sets contains 40 events, 40 generated outputs, and no missing outputs. Here N counts utterance IDs, while 24 denotes the four language pairs crossed with six policies; the bootstrap unit is the paired utterance ID, not a speaker. Appendix Table 5 maps these counts and the relation between the pilot and WavLM follow-up. For each predeclared control, we compute source-policy minus control deltas paired by identifier. The pooled result uses 10,000 percentile-bootstrap resamples of the 160 aligned deltas with seed 20260918; language-pair intervals use the same procedure within each 40-item language pair. We report descriptive intervals without multiplicity adjustment and do not use them as a family-wise sequence of hypothesis tests. The pooled summaries are utterance-conditional aggregates; when language-pair signs differ,

the corresponding tables and prose carry the main interpretation. Intervals that remain compatible with both signs are inconclusive at this paired sample size, not evidence of equivalence. Because speaker_id is unavailable in the current record, the bootstrap does not establish speaker-level independence. These intervals cover sampled utterances, not speaker, document, prompt-bank, language, or listener populations.

WavLM follow-up protocol and cohort relation. A recorded, support-screened FLEURS index contains 20 items per language pair ($N = 80$) under the same six policies, fixed text, renderer, candidate bank, and decoding settings. A stable-ID comparison finds 29 follow-up items also present in the 160-item pilot (5, 8, 11, and 5 by language pair); the other 51 follow-up IDs do not occur in the pilot. The follow-up is therefore not an independent-cohort replication. Here support-screened means that each reported comparison uses paired utterance IDs with valid outputs and finite scores in both conditions. The follow-up protocol and per-pair assignments were recorded in private, version-controlled provenance before follow-up synthesis; this is an internal record, not third-party preregistration, and the record does not establish that the overlapping items' earlier pilot outcomes were unseen. All 24 language-pair-by-policy output sets contain 20 events and zero missing waveforms. Frozen WavLM Base+ uses the same resolved revision as the pilot; source-minus-control intervals use 10,000 paired-utterance bootstrap resamples with seed 20260919. The intervals are conditional on these 80 selected IDs and are not independent of the pilot; they do not support speaker-, language-, or S2ST-population inference. The primary comparison is source acoustic minus function-label-only; shuffled-source profiles are the second predeclared control. The positive rule requires both source-minus-function-label and source-minus-shuffled-source intervals to be above zero; it is not met.

4 Results: selector-aligned and target-side readouts diverge under controls

Table 1 reports pooled contrasts, and Table 2 shows the function-label comparison within each language pair. Source-acoustic prompting has a higher selector-aligned source-summary score than the recorded shuffled-source allocation (+0.0716, 95% interval $[+0.0326, +0.1128]$). This is one descriptive interval among multiple unadjusted contrasts, on a score that shares the selector's feature family; it does not show that generated delivery matches a target reference.

Against function-label-only prompting, the source-summary difference is -0.0197 with an interval compatible with zero. All three random-policy contrasts are also compatible with zero in this fixed sample. Table 2 shows why the pooled target-side summaries need caution: the target-reference estimate is below zero in every function-label pair, whereas the target-duration estimate is positive in English–German ($+0.0274$ $[+0.0087, +0.0419]$) and negative in the other three pairs. The pooled target-reference summary is lower for source-acoustic prompting against function-label-only (-0.1606) and the three random policies (-0.0915, -0.1102, and -0.1211); those intervals are below zero. The pooled target-duration readout is lower by -0.1370 against function labels and by -0.0531, -0.0594, and -0.0522 against the three random policies, but these pooled values describe utterance-level aggregates and do not establish stability across language pairs. Target-duration absolute error is $|d_{\text{gen}} - d_{\text{target}}|$; its pooled mean is 5.3575 seconds for source-acoustic prompting and 3.8689 seconds for function-label-only, a paired difference of +1.4886 seconds (95% interval $[+1.0960, +1.8772]$ seconds). We report this seconds-scale quantity in text because Table 1 reports normalized similarities. The held-out reference durations have mean 11.1270 seconds and median 10.7100 seconds in this sample. The shuffled-source contrasts are $+0.0043$ $[-0.0334, +0.0441]$ for target reference and $+0.0038$ $[-0.0293, +0.0374]$ for duration, both compatible with both signs.

Figure 2 shows the control pattern on a common difference scale. Of the displayed contrasts, only the selector-aligned source-summary comparison with the recorded shuffled-source allocation has a positive descriptive interval that excludes zero. The target-reference and target-duration panels do not show the same pattern, so the figure remains a comparison of automatic readouts, not delivery evidence.

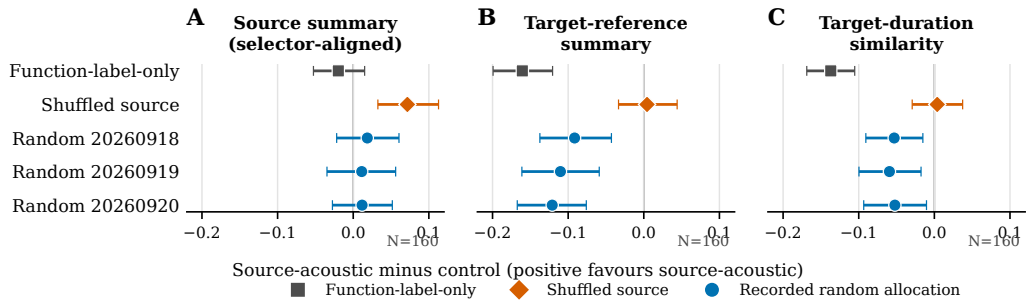


Figure 2: Paired source-acoustic minus control differences on three automatic readouts, pooled over 160 fixed-text utterances. Horizontal lines are 95% percentile-bootstrap intervals from 10,000 utterance-level resamples. Positive values favour source-acoustic prompting. The source-summary panel is selector-aligned; the remaining panels are automatic target-side diagnostics and are not listener outcomes. Random policies A/B/C use allocation seeds 20260918/20260919/20260920, respectively.

Table 1: Paired automatic contrasts pooled over 160 fixed-text utterances. Every entry is source-acoustic prompting minus the named control; brackets are 95% percentile-bootstrap intervals from 10,000 utterance-level resamples. Larger values are better. Target duration uses only waveform duration and excludes the selector fields by construction, but remains an automatic timing diagnostic.

Control	N	Source-summary Δ ↑	Target-reference Δ ↑	Target-duration Δ ↑
Function-label-only	160	-0.0197 $[-0.0526, +0.0151]$	-0.1606 $[-0.1993, -0.1206]$	-0.1370 $[-0.1687, -0.1053]$
Shuffled source profile	160	$+0.0716$ $[+0.0326, +0.1128]$	$+0.0043$ $[-0.0334, +0.0441]$	$+0.0038$ $[-0.0293, +0.0374]$
Random policy A	160	$+0.0186$ $[-0.0219, +0.0605]$	-0.0915 $[-0.1374, -0.0428]$	-0.0531 $[-0.0906, -0.0152]$
Random policy B	160	$+0.0112$ $[-0.0347, +0.0561]$	-0.1102 $[-0.1612, -0.0589]$	-0.0594 $[-0.0997, -0.0176]$
Random policy C	160	$+0.0117$ $[-0.0275, +0.0517]$	-0.1211 $[-0.1673, -0.0760]$	-0.0522 $[-0.0934, -0.0104]$

The pooled result hides heterogeneity. Table 2 gives the within-language-pair intervals for the function-label contrast. The source summary differs in sign across pairs, and the target-reference point estimate is below zero in each pair, but the English–Dutch interval remains compatible with both signs. In contrast, the duration estimate is positive for English–German ($+0.0274$ $[+0.0087, +0.0419]$) and negative for the other three pairs. Its corresponding absolute-error contrast is negative for English–German and positive for the other three pairs. Therefore neither pooled duration statistic supports a cross-language-pair timing statement. These are not independent language replications and do not support a language-general result.

Stricter target-frame and learned-representation audits. Table 3 reports the stricter summary-exclusion audit, which excludes the selector's four stored scalar fields from score inputs, derives residual pitch, energy, and pause channels, and keeps duration separate. A separate seven-case controlled validator reproduces the intended responses on synthetic perturbations and retains stored-summary-only and nuisance-only invariance. This supports the scorer's stated access pattern; it does not establish real-data coverage or an independent delivery endpoint. Against function labels, source-acoustic prompting is lower on pooled energy projection (-0.0760 $[-0.1224, -0.0315]$) and duration (-0.1370 $[-0.1678, -0.1062]$), while pitch projection (-0.0313 $[-0.0937, +0.0304]$) and pause projection (-0.0109 $[-0.0482, +0.0248]$) remain compatible with both signs. Against shuffled source profiles, only pitch projection is positive ($+0.0732$ $[+0.0051, +0.1379]$); energy, duration, and pause remain compatible with both signs. Because the paper reports many control-by-metric contrasts without multiplicity adjustment, we treat that isolated positive shuffled-source pitch interval as a residual automatic diagnostic, not standalone positive evidence. Its 72 paired identifiers comprise 40 English–German, 27 English–French,

Table 2: Within-pair source-acoustic minus function-label-only contrasts. Each interval is a 10,000-replicate utterance bootstrap over 40 fixed items. Larger values are better.

Pair	Source-summary Δ ↑	Target-reference Δ ↑	Target-duration Δ ↑
English–German	$+0.0515$ $[+0.0142, +0.0898]$	-0.0615 $[-0.0979, -0.0259]$	$+0.0274$ $[+0.0087, +0.0419]$
English–French	$+0.0622$ $[+0.0031, +0.1265]$	-0.1829 $[-0.2653, -0.0995]$	-0.1068 $[-0.1350, -0.0804]$
English–Dutch	-0.0446 $[-0.1229, +0.0500]$	-0.0778 $[-0.1609, +0.0180]$	-0.2743 $[-0.3403, -0.2055]$
English–Turkish	-0.1478 $[-0.1969, -0.0978]$	-0.3204 $[-0.3908, -0.2469]$	-0.1944 $[-0.2645, -0.1184]$

Table 3: Strict summary-exclusion target-frame contrasts pooled by available paired identifiers. Every entry is source-acoustic prompting minus the named control; brackets are 95% percentile-bootstrap intervals from 10,000 utterance-level resamples. Pitch, energy, and pause use residual channels that exclude the selector's stored scalar fields from score inputs; duration remains a separate waveform-duration statistic. Each metric uses only identifiers finite in both compared conditions; the panel note reports the pooled denominators.

Control	Pitch Δ ↑	Energy Δ ↑
Function-label-only	-0.031 $[-0.094, +0.030]$	-0.076 $[-0.122, -0.032]$
Shuffled source profile	$+0.073$ $[+0.005, +0.138]$	0.000 $[-0.052, +0.050]$
Random policy A	$+0.008$ $[-0.059, +0.073]$	-0.020 $[-0.072, +0.033]$
Random policy B	$+0.011$ $[-0.054, +0.074]$	-0.051 $[-0.108, +0.005]$
Random policy C	$+0.005$ $[-0.059, +0.068]$	-0.027 $[-0.075, +0.021]$

Control	Duration Δ ↑	Pause Δ ↑
Function-label-only	-0.137 $[-0.168, -0.106]$	-0.011 $[-0.048, +0.025]$
Shuffled source profile	$+0.004$ $[-0.030, +0.037]$	$+0.009$ $[-0.036, +0.055]$
Random policy A	-0.053 $[-0.092, -0.016]$	-0.001 $[-0.049, +0.046]$
Random policy B	-0.059 $[-0.100, -0.018]$	-0.002 $[-0.038, +0.034]$
Random policy C	-0.052 $[-0.094, -0.010]$	-0.016 $[-0.065, +0.032]$

Pooled paired identifiers (pitch/energy/duration/pause): function-label 84/84/160/75; shuffled source 72/71/160/60; random A 79/79/160/69; random B 78/77/160/68; random C 80/80/160/71.

no English–Dutch, and five English–Turkish items; every supported pair-level interval includes zero. The pooled interval therefore does not describe a four-pair effect. The pooled denominators are 84/84/160/75 against function labels and 72/71/160/60 against shuffled source for pitch, energy, duration, and pause, while the random controls range from 77 to 80 for pitch and energy and from 68 to 71 for pause. English–Dutch still has unsupported shuffled-source and random-policy-B contrasts for pitch, energy, and pause because the projected-metric support screen leaves no finite paired identifiers there. The earlier field-masked audit in Appendix A points in the same direction and attributes its missing pitch and energy coverage to limited support in the held-out references, not missing generation. A separate summary-invariant residual-contour audit there removes a direct pause-count path but does not change the target-side conclusion. Frozen WavLM target-reference similarity is lower against function labels (-0.0690 $[-0.0869, -0.0519]$) and all three random allocations (-0.0319 to -0.0347), while its shuffled-source contrast is $+0.0002$ $[-0.0153, +0.0159]$. The support-screened WavLM follow-up in Table 4, with its protocol and assignments recorded in private run provenance before synthesis, repeats the lower function-label contrast (-0.0418 $[-0.0615, -0.0232]$) and has an inconclusive shuffled-source contrast ($+0.0013$ $[-0.0180, +0.0206]$). Table 7 shows that the function-label point estimate is negative in all four pairs, but only English–Turkish excludes zero, so the pooled negative contrast is concentrated there and is not evenly shared across pairs. This is another automatic readout. Together, these checks do not confirm a target-side benefit for the selector policy and do not establish an independent replication.

Interpretation. The recorded shuffled allocation checks whether the configured source profile affects the aligned summary through this selector. It does not make that summary feature-disjoint, establish robustness to other shuffles, or identify a source-side communica-

Table 4: Frozen-protocol WavLM results on 80 support-screened follow-up FLEURS items, 29 of which overlap with the pilot. Each value is source-acoustic prompting minus the named control; brackets are 95% percentile-bootstrap intervals from 10,000 paired utterance resamples conditional on these selected IDs. Larger values are better. All 480 generated waveforms are present. The primary function-label comparison is below zero, and the shuffled-source control is compatible with both signs; the positive rule is not met.

Control	N	Frozen WavLM Δ ↑
Function-label-only	80	-0.0418 $[-0.0615, -0.0232]$
Shuffled source profile	80	$+0.0013$ $[-0.0180, +0.0206]$
Random policy A	80	-0.0294 $[-0.0505, -0.0098]$
Random policy B	80	-0.0058 $[-0.0215, +0.0101]$
Random policy C	80	-0.0173 $[-0.0359, +0.0004]$

tive property. The target-reference differences against function labels and random allocations are automatic diagnostics under the stated renderer, candidate bank, 40-item subsets, and fixed allocations. The duration diagnostic excludes the selector fields, but its function-label effect reverses in one pair. It remains one timing statistic whose physical correlation with pause structure is not removed by that exclusion. The broader point is methodological: paired controls can show when an apparently favorable automatic score is tied to the selector but does not correspond to consistent target-side behavior. These results do not rank source-conditioned methods or establish listener-level harm.

5 Implications for prompt-policy evaluation

The readouts answer different questions and should not be collapsed into a single measure of delivery quality. The selector-aligned score tests whether an output moves toward the feature family used to choose its prompt. The target-reference score instead compares generated speech with held-out speech for the same translated item, while duration and WavLM provide distinct automatic checks with their own limits. Here, movement on the aligned score against one shuffled allocation is not matched by a target-reference or WavLM benefit; the comparisons with function-label-only and random policies are inconclusive or lower. The strict summary-exclusion pitch contrast also depends on finite support: its pooled positive interval uses 72 paired IDs, includes no English–Dutch items, and none of the supported language-pair intervals excludes zero (Table 3). The denominator composition therefore changes how that pooled result should be read. Future prompt audits should predefine a selector-independent target endpoint and its support rules, preserve paired controls, and report finite denominators by language and metric. This keeps metric alignment distinct from target matching and leaves listener-level conclusions to listener evidence.

6 Limitations

This study has ten central limits. First, the source summary used by the evaluated source-acoustic selector shares a feature family with that selector, which is the coupling this audit is designed to expose. Second, the target-reference score is still an automatic acoustic diagnostic. Third, the duration diagnostic excludes selector fields, but its function-label contrast reverses in one pair, can covary with pause structure, and therefore does not provide a stable perceptual endpoint. Fourth, the stricter summary-exclusion audit excludes the selector's stored scalar fields from score inputs, and its narrow controlled validator shows the scorer reacting as intended on synthetic perturbations, but the real paired audit still has unsupported English–Dutch shuffled-source and random-policy-B pitch, energy, and pause contrasts, so it does not provide a complete feature-disjoint confirmation set. The appendix-only summary-invariant audit also excludes those scalar inputs but remains frame-derived and does not change the target-side conclusion. Fifth, frozen WavLM is a separate learned automatic representation, but not listener evidence or a feature-family-disjoint endpoint; its follow-up does not meet the positive rule, and its pooled negative function-label contrast is concentrated in English–Turkish. Because 29 of 80 follow-up item IDs overlap the pilot, it is

not an independent replication. Sixth, both item sets use read-speech validation data, and the current records contain no speaker_id, so speaker overlap and speaker-disjoint generalization cannot be assessed. Seventh, the paper does not include an oracle or upper-bound prompt condition, so score magnitudes should be interpreted alongside paired controls; without an oracle condition, they do not estimate headroom. Eighth, the fixed-text statement concerns the TTS input string, not verified lexical identity of the generated speech; we do not add a separate lexical-fidelity audit across prompt policies. Ninth, no listener study, perceptual annotation, or deployment evaluation was performed. Tenth, the staged anonymous code package does not redistribute private audio or the full frozen run record. The results do not establish pragmatic-function preservation, naturalness, preference, speaker transfer, or general S2ST quality.

The next necessary step is a broader feature-disjoint evaluation. The completed strict summary-exclusion audit is the strongest current stored-summary-exclusion check, but it still leaves unsupported Dutch shuffled/random contrasts and therefore cannot serve as a complete primary endpoint. The earlier field-masked reanalysis likewise remains related to selector fields through raw frames. The completed WavLM follow-up adds a separate learned automatic readout, but not separate pitch, energy, rate, and pause measurements with verified exclusion from retrieval. Any future feature-disjoint study must predeclare complete coverage, retain the same paired policy comparisons, and validate controlled perturbations before synthesis. Listener evaluation would remain a separate requirement.

7 Conclusion

Fixed-text counterfactuals make a narrow S2ST question testable: what changes when only prompt policy changes? In this four-language pilot, one source-acoustic selector moves its aligned automatic summary relative to one shuffled allocation, while its contrast with function-label-only and random allocations is inconclusive. Its target-reference diagnostic and frozen WavLM similarity are lower against function-label-only and random allocations but inconclusive against the shuffled allocation. The pooled duration diagnostic follows that direction, yet reverses in English–German and cannot support a cross-pair timing result. A stricter summary-exclusion scorer passes a narrow controlled validation, but its paired audit still only recovers a positive shuffled-source contrast on pitch projection; energy, duration, and pause remain inconclusive or negative, and some Dutch shuffled/random contrasts have no paired support. In the support-screened 80-item WavLM follow-up, 29 item IDs overlap with the pilot; its complete-coverage similarity is lower against function labels and inconclusive against shuffled source profiles, with the negative contrast concentrated in English–Turkish. This follow-up does not meet the positive rule and is not an independent replication. The main contribution is therefore methodological: a paired fixed-text audit can reveal when a prompt policy improves a selector-aligned signal without producing a consistent target-side gain. A broader feature-disjoint evaluation, an oracle upper-bound condition, a lexical-fidelity check, and separately authorized listener evidence are needed before extending this diagnostic to source-to-target delivery transfer.

AI Use

We used AI tools to find related work and to help edit and polish the paper's writing.

Reproducibility Statement

Section 3 describes the prompt policies and automatic scores; Appendix A gives their definitions, filtering rules, and paired-uncertainty procedure. The supplementary source ZIP contains the evaluation and reporting code and tests under the MIT License. It excludes audio, model weights, and per-item run records, so it cannot reproduce the reported scores by itself; dataset and model terms remain those of their providers.

References

Yanjie An, Yuxiang Zhao, Yichi Zhang, Qixi Zheng, Yujie Tu, Keqi Deng, Kai Yu, and Xie Chen. Openstbench: Beyond semantic evaluation for speech translation, 2026. URL https://arxiv.org/abs/2605.30792. arXiv.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech, 2022. URL https://arxiv.org/abs/2205.12446. arXiv:2205.12446.

Feng Jiang, Zhiyu Lin, Yiyang Liu, Liumeng Xue, Fan Bu, Yuhao Du, Xiangying Chen, Benyou Wang, and Haizhou Li. S2s-arena: Evaluating paralinguistic instruction following in speech-to-speech models, 2026. URL https://aclanthology.org/2026.acl-long.1615/. ACL 2026.

Alkis Koudounas, Hayato Futami, Quentin Jodelet, Osamu Take, Shinji Watanabe, and Emiru Tsunoo. Benchmarking speech-to-speech translation models, 2026. URL https://arxiv.org/abs/2606.03241. arXiv.

Seamless Communication et al. Seamless: Multilingual expressive and streaming speech translation, 2023a. URL https://arxiv.org/abs/2312.05187. arXiv.

Seamless Communication et al. Seamlessm4t: Massively multilingual and multimodal machine translation, 2023b. URL https://arxiv.org/abs/2308.11596. arXiv.

Ismail Rasim Ulgen, Nancy Liu, Najmeh Sadoughi, Abhishek Yanamandra, Abhinav Jain, Zhu Liu, and Vimal Bhat. Direct transfer of prosody in speech-to-speech translation using disentangled speech tokens, 2026. URL https://ieeexplore.ieee.org/document/11464610. IEEE Xplore 2026.

Nigel G. Ward, Marcel de Korte, Javier Vazquez, Montserrat G. Molina, Vanessa Bolado, Carol Figueroa, Eliya Nachmani, John E. Ortega, and Satoshi Nakamura. The interspeech 2026 challenge on transfer of pragmatic intent in speech-to-speech translation, 2026. URL https://www.cs.utep.edu/topi/. Interspeech 2026 Challenge.

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. URL https://www.isca-archive.org/interspeech_2024/xue24b_interspeech.html.

A Evidence map and protocol details

This appendix maps the study samples and units before describing the automatic readouts that support the main paper. Table 5 distinguishes unique utterance IDs, policy-by-language result cells, condition-specific events, and analysis units. The sections below then detail the selector-aligned and duration diagnostics, the field-masked, summary-exclusion, and learned-representation analyses, and the scope limits that keep all of these measures separate from listener evidence.

Table 5: Study identity and count key. N counts unique utterance IDs; a result cell is one language pair crossed with one prompt policy; an event is one condition-specific output record.

Evidence stage	Item sample	Policy and output map	Relation and analysis unit
Pilot	160 IDs; 40 in each of four language pairs	Six policies crossed with four pairs give 24 result cells; 40 events per cell	The same IDs are compared across policies. Bootstrap resamples paired utterance IDs: 160 pooled or 40 within a pair.
Pilot reanalyses	Reuse the pilot IDs; no new cohort	Field-masked, summary-exclusion, summary-invariant, and frozen-WavLM readouts; finite paired counts vary by metric	Reanalyses are distinct score families on the pilot, not independent samples. Each metric uses its stated paired denominator.
WavLM follow-up	80 IDs; 20 in each pair; 29 overlap pilot (5, 8, 11, 5), 51 do not	The same six policies give 24 result cells; 20 events per cell	The follow-up is partially overlapping, not an independent replication. Bootstrap resamples paired utterance IDs conditional on these selected items. Speaker overlap is unknown.
Synthetic scorer check	Seven controlled cases; no natural-speech cohort	Designated pitch, energy, duration, and pause changes plus invariance checks	This checks scorer behavior only; it is not a sample-level endpoint.

The six policies are source-acoustic, function-label-only, one shuffled-source allocation, and three recorded random allocations. No inspected cohort record contains a speaker ID, so utterance-level resampling does not establish speaker-level independence.

Selector-aligned source summary. Under the recorded shuffled-source allocation, source-acoustic prompting has a higher configured source-summary score ($+0.0716$ $[+0.0326, +0.1128]$). Because this score shares the selector's feature family, it is not independent transfer evidence.

Target-duration diagnostic. The pooled source-prompting contrast is lower against function labels (-0.1370 $[-0.1687, -0.1053]$) and each recorded random allocation (-0.0531, -0.0594, and -0.0522); the shuffled-source contrast is $+0.0038$ $[-0.0293, +0.0374]$ and compatible with both signs. The function-label duration contrast reverses in English–German ($+0.0274$ $[+0.0087, +0.0419]$), so the pooled number is not a cross-language-pair timing result. Target-duration absolute error is $|d_{\text{gen}} - d_{\text{target}}|$: its pooled mean is 5.3575 seconds for source prompting and 3.8689 seconds for function-label-only, a paired difference of $+1.4886$ seconds (95% interval $[+1.0960, +1.8772]$ seconds); held-out reference durations have mean 11.1270 and median 10.7100 seconds. Table 6 exposes the corresponding language-pair physical-unit results. This readout uses waveform duration only. An implementation unit test builds a 1-second sine reference and separate 1-second pitch, amplitude, internal-silence-replacement, and stereo variants; all retain 1-second duration and give similarity 1, while a 0.5-second waveform gives 0.5. This demonstrates the score's direct-field behavior, not that duration is a valid delivery endpoint. Duration can still covary with pauses in natural speech.

Table 6: Within-language-pair target-duration absolute error in seconds for source-acoustic and function-label-only prompting. The final column is source-acoustic minus function-label-only with a 95% percentile-bootstrap interval over 40 paired utterances.

Pair	Source-acoustic mean (s)	Function-label mean (s)	Paired Δ (s, 95% interval)
English–German	4.8076	5.0727	-0.2651 $[-0.4698, -0.0112]$
English–French	3.4066	2.3042	$+1.1024$ $[+0.7901, +1.4351]$
English–Dutch	4.9467	2.3389	$+2.6078$ $[+1.8347, +3.3971]$
English–Turkish	8.2690	5.7599	$+2.5091$ $[+1.3924, +3.5519]$

Field-masked target-frame diagnostics. The field-masked reanalysis re-scores all 24 completed condition files from frozen target WAVs. It excludes the selector's stored mean-pitch, mean-RMS, voiced-ratio, and pause-count fields, then compares centered time-normalized log-pitch shape, log-RMS shape outside pauses, pause-boundary F1, and duration. All files have 40 generated outputs and zero output missingness. Pitch and energy require enough suitable frames, so their paired denominators are 119/119 against function labels, 111/112 against shuffled source profiles, and 113–116 against random allocations; pause uses 160 items throughout. Against function labels, source-acoustic prompting is lower by -0.0284 $[-0.0542, -0.0028]$ for pitch shape and -0.0349 $[-0.0554, -0.0153]$ for energy shape; pause-boundary F1 is -0.0097 $[-0.0262, +0.0071]$. Every shuffled-source interval includes zero. A read-only availability audit attributes 28 of the 40 missing English–Dutch pitch scores, and likewise 28 of the 40 missing energy scores, to held-out reference support; none come from insufficient generated frames in that condition. Synthetic profile tests isolate the designated pitch, energy, pause, and duration changes, but raw frame values can still encode the excluded summaries. This is therefore a field-masked automatic audit with metric-specific paired denominators, not a feature-disjoint delivery endpoint.

Strict summary-exclusion target-frame diagnostics. A stricter reanalysis does not read the selector's four stored scalar fields—mean pitch, mean RMS energy, voiced-frame ratio, or pause count—when constructing score inputs. It derives centered, resampled, DC-excluded residual vectors for log pitch, log RMS outside pauses, and pause occupancy, compares target and generated vectors with rescaled cosine similarity, and scores duration separately. Against function labels, pooled source-minus-control contrasts are -0.0313 $[-0.0937, +0.0304]$ for pitch projection, -0.0760 $[-0.1224, -0.0315]$ for energy projection, -0.1370 $[-0.1678, -0.1062]$ for duration, and -0.0109 $[-0.0482, +0.0248]$ for pause projection. Against shuffled source profiles, the corresponding contrasts are $+0.0732$ $[+0.0051, +0.1379]$, -0.0003 $[-0.0517, +0.0503]$, $+0.0038$ $[-0.0302, +0.0367]$, and $+0.0090$ $[-0.0362, +0.0549]$. The pooled denominators are 84/84/160/75 against function labels, 72/71/160/60 against shuffled source, 79/79/160/69 against random policy A, 78/77/160/68 against random policy B, and 80/80/160/71 against random policy C for pitch, energy, duration, and pause. English–Dutch has no finite paired pitch, energy, or pause values against shuffled source or random policy B, so those pair-level contrasts are marked unsupported, and no values are imputed. This is a stronger stored-summary-exclusion audit than the field-masked analysis, but it still does not establish a complete feature-disjoint endpoint.

Summary-invariant target-frame diagnostics. A separate read-only analysis standardizes full log-pitch and log-RMS contours and compares them by correlation, retains waveform duration, and uses maximum pause fraction instead of the strict scorer's projected pause occupancy. It does not read mean pitch, mean RMS energy, voiced-frame ratio, or pause count when scoring. The residual pitch and energy values agree with the field-masked values to the displayed precision because correlation is already invariant to affine shifts and scales. Maximum-pause-fraction source-minus-control contrasts are -0.0382 $[-0.0866, +0.0113]$ against function labels, $+0.0249$ $[-0.0194, +0.0708]$ against shuffled source profiles, and -0.0717 $[-0.1285, -0.0177]$ against random seed 20260920; all use 160 paired identifiers. The remaining random intervals include zero. This removes a direct pause-count path but still derives temporal features from speech frames, so it remains an automatic summary-invariant audit that does not provide an independent endpoint.

Frozen WavLM target-reference diagnostic. A frozen-output reanalysis uses mean-pooled embeddings from frozen microsoft/wavlm-base-plus, resolved to revision 4c66d4806a428f2e922ccfa1a962776e232d487b, on held-out target references and generated target WAVs. The model is not available to prompt selection. Source-acoustic prompting minus function-label-only is -0.0690 $[-0.0869, -0.0519]$; the three random contrasts are -0.0335 $[-0.0501, -0.0173]$, -0.0319 $[-0.0508, -0.0130]$, and -0.0347 $[-0.0511, -0.0184]$; the shuffled-source contrast is $+0.0002$ $[-0.0153, +0.0159]$. All contrasts use 160 paired identifiers and 10,000 percentile-bootstrap resamples. This separate pretrained model architecture remains an automatic diagnostic that does not provide an independent endpoint.

WavLM follow-up. The support-screened FLEURS index has 20 items per pair and 80 IDs overall, of which 29 overlap with the pilot (5, 8, 11, and 5 by pair); 51 do not occur in the pilot. The protocol and per-pair assignments were recorded in private, version-controlled run provenance before follow-up synthesis. Because some IDs were already used in the pilot, this ordering does not establish an independent cohort or that earlier outcomes for the shared items were unseen; the result is a protocolized follow-up, not an independent replication. All 24 language-pair-by-policy outputs contain 20 events and zero missing waveforms. With the same frozen WavLM revision and 10,000 paired bootstrap resamples (seed 20260919), source acoustic minus function-label-only is -0.0418 $[-0.0615, -0.0232]$ and source acoustic minus shuffled source is $+0.0013$ $[-0.0180, +0.0206]$. These intervals are conditional on the selected 80 IDs and do not provide speaker-level inference. Table 7 shows the two predeclared contrasts by pair. The function-label point estimate is negative in all four pairs, but only English–Turkish excludes zero, while every shuffled-source interval contains zero. The primary and shuffled-source-control intervals therefore do not satisfy the positive automatic rule. This is an automatic readout on a partially overlapping follow-up sample, not independent validation.

Table 7: Frozen-protocol WavLM follow-up by language pair. Each entry is source-acoustic prompting minus the named control with a 95% percentile-bootstrap interval over 20 paired items. The follow-up overlaps the pilot on 29 IDs, so these results are not an independent replication. Larger values are better.

Pair	Function-label-only Δ ↑	Shuffled-source Δ ↑
English–German	-0.0039 $[-0.0082, +0.0001]$	$+0.0011$ $[-0.0055, +0.0085]$
English–French	-0.0048 $[-0.0103, +0.0009]$	-0.0013 $[-0.0044, +0.0018]$
English–Dutch	-0.0363 $[-0.0799, +0.0031]$	$+0.0025$ $[-0.0440, +0.0478]$
English–Turkish	-0.1222 $[-0.1699, -0.0751]$	$+0.0028$ $[-0.0589, +0.0663]$

Every paired contrast uses the same utterance identifier in both conditions. The analysis rejects a condition when its evaluation file reports missing generated outputs, has duplicate identifiers, or has a segment-count mismatch. The report uses 10,000 percentile bootstrap resamples of aligned utterance-level deltas with fixed seed 20260919. Pooled intervals resample utterances and therefore do not identify language-, speaker-, prompt-bank-, or document-level variability.

B What the protocol does not evaluate

No listener, deployment, or speaker-disjoint evaluation was run. The target-reference and duration scores are automatic acoustic diagnostics; duration can covary with pauses. Field-masked scores retain raw pitch and RMS, while summary-exclusion and summary-invariant scores remain frame- or pause-derived. The strict scorer has unsupported Dutch shuffled/random contrasts. Its seven synthetic cases check responsiveness and invariance, not real-data coverage or listener validity. The selector-aligned source summary shares a feature family with the selector. A broader test would need a feature-disjoint evaluator, complete paired coverage, and controlled pitch, energy, rate, pause, and channel perturbations.