

Early Warning for Target-Speaker Extraction: Reliability under Distribution and Failure-Definition Shift

Shubham Sharma and Manik Gupta

Department of Computer Science and Information Systems
Birla Institute of Technology and Science Pilani, Hyderabad Campus
Hyderabad, India
{p20240435,manik}@hyderabad.bits-pilani.ac.in

Abstract

Target-speaker extraction (TSE) systems are typically evaluated only after the complete output has been produced, limiting the ability to identify potentially unreliable outputs during generation. We investigate whether failure risk can instead be estimated causally from only the first three seconds of the emitted TSE waveform.

We develop an early-warning monitor using 11,946 outputs spanning five conversational corpora and six TSE systems. Because independent failure labels are unavailable for this discovery cohort, we define a proxy adverse outcome from seven complete-output indicators of degraded or atypical TSE behaviour. After excluding monitoring features that overlap with the proxy construction, the monitor predicts this eventual outcome from the early waveform with an AUROC of 0.8062 and an AUPRC of 0.6606. At a 15% review budget, it identifies 54.47% of proxy-adverse outputs with 66.29% precision.

To test whether this early-warning signal transfers beyond the discovery setting, we freeze the monitor and evaluate it unchanged on 2,000 unseen outputs with independently computed SI-SDRi. When failure is defined as $\text{SI-SDRi} \leq 0$ dB, performance falls to near chance (AUROC 0.5354; AUPRC 0.0700). Yet, on the same external outputs, the frozen predictions remain strongly associated with the original proxy outcome (AUROC 0.8092; AUPRC 0.8481). The proxy and SI-SDRi failure definitions agree only weakly ($\kappa = 0.028$, $\text{MCC} = 0.064$), while substantial feature-distribution shift is observed between the discovery and external cohorts (45 of 63 features have $\text{PSI} \geq 0.25$).

These results show that short-prefix monitoring can provide useful early risk prioritization within the setting in which the monitor is developed, but that such signals do not automatically transfer to independently defined failure outcomes. Reliable deployment therefore

requires explicit attention to both distribution shift and the definition of failure.

Keywords: target-speaker extraction; reliability monitoring; early failure detection; distribution shift; out-of-domain evaluation; selective prediction.

1 Introduction

Target-speaker extraction (TSE) aims to recover the speech of a designated speaker from a multi-speaker mixture using an enrollment utterance or speaker representation as the target cue (Delcroix et al., 2018; Zhang et al., 2025). Although modern TSE systems perform well on controlled benchmarks, assessing their reliability is more difficult in real conversational audio, where overlapping speech, reverberation, noise, intermittent target activity, and enrollment–mixture mismatch are common. Under such conditions, a TSE system may suppress the intended speaker, retain an interfering speaker, or distort the temporal structure of the extracted speech.

Existing evaluation protocols largely assess TSE outputs after extraction is complete, using aggregate measures of intelligibility, speaker similarity, timing, and perceptual quality. These measures are valuable for comparing systems, but they do not address an important runtime question: can we identify an output that is likely to become unreliable while it is still being produced? An early causal monitor could provide such a warning and support abstention, escalation, or downstream review without modifying the extraction waveform itself.

Developing such a monitor raises two main challenges. The first is how to define an unreliable output. Independent failure labels are not available for all outputs in our discovery cohort. We therefore construct a proxy adverse label from seven complete-output indicators of degraded or

atypical TSE behaviour: silence, clipping, energy mismatch, spectral flatness, temporal instability, output–mixture inconsistency, and cross-system disagreement. Because some candidate monitoring features may overlap with the indicators used to construct this proxy, prediction could be artificially inflated by target circularity. We therefore explicitly test whether useful early prediction remains after removing these overlapping features.

The second challenge is whether a monitor developed in one setting remains reliable in another. Its performance may change across conversational corpora or TSE architectures, particularly when the underlying data distribution changes. More fundamentally, performance may also depend on what is considered a failure. We therefore examine the monitoring approach across multiple corpora and TSE systems and subsequently test a frozen monitor on unseen outputs using an independently computed reconstruction-based definition of failure.

Our central question is therefore: *can the first three seconds of a TSE output provide an early warning that the completed output will be labelled as adverse?* We investigate this using a single-stream causal monitor that observes only the waveform emitted by the evaluated TSE system. The monitor has no access to future audio, does not generate or compare alternative outputs, and does not modify or replace the emitted waveform. It simply assigns a risk score from the acoustic evidence available in the early output. We then examine how well this risk ranking generalizes across corpora, TSE systems, and failure definitions.

We make four main contributions. First, we formulate TSE reliability monitoring as a single-stream causal risk-prediction problem that uses only the first three seconds of one emitted waveform. Second, we explicitly audit target circularity and show that a predefined reduced non-overlap feature set retains substantial predictive signal under trial-disjoint evaluation, achieving an AUROC of 0.8062 and an AUPRC of 0.6606. Third, we evaluate the practical utility of the resulting risk ranking under limited review budgets and examine how consistently the same policy performs across five conversational corpora and six TSE systems.

Fourth, we freeze the trained monitor and apply it unchanged to 2,000 unseen outputs from four TSE systems. These outputs provide an independently computed reconstruction measure, SI-SDRi, which indicates whether TSE improves

the target signal relative to the original mixture. When failure is defined as $\text{SI-SDRi} \leq 0$ dB, the frozen monitor performs close to chance (AUROC 0.5354; AUPRC 0.0700). Yet, on these same external outputs, its predictions remain strong against the original proxy-defined outcome (AUROC 0.8092; AUPRC 0.8481). Together with substantial differences between the discovery and external feature distributions and weak agreement between the two failure labels, these findings show that external reliability depends not only on how the data change, but also on how failure is defined.

Overall, our results support causal early warning as a useful mechanism for prioritizing potentially unreliable TSE outputs within the setting in which the monitor is developed. At the same time, they show that strong within-setting performance should not be interpreted as evidence of universal failure detection: reliable transport to new data therefore requires explicit attention to both distribution shift and independently grounded definitions of failure.

2 Related Work

Target-speaker extraction. Target-speaker extraction conditions a separation model on an enrollment utterance or speaker representation to recover a designated speaker from a mixture. SpeakerBeam established enrollment-conditioned extraction for single-channel speech and was later extended to target-speaker recognition (Delcroix et al., 2018, 2019). Subsequent work strengthened speaker conditioning through improved embedding architectures and multi-level speaker representations (Desplanques et al., 2020; Wang et al., 2023; Zhang et al., 2025). Other methods explicitly address speaker-confusion errors arising from ambiguous target-speaker representations (Zhao et al., 2022; Liu et al., 2023), while band-split recurrent modelling provides an effective time–frequency reconstruction architecture for source separation (Luo and Yu, 2023). These studies primarily improve or repair the extraction process itself. In contrast, we investigate whether an output produced by a fixed TSE system can be identified as high risk from early causal evidence, without generating alternatives or modifying the emitted waveform.

Confidence estimation and selective prediction. Selective prediction allows a model to abstain or defer when its estimated risk is high, inducing

an explicit trade-off between prediction coverage and reliability (Geifman and El-Yaniv, 2019). In speech recognition, confidence-estimation methods identify potentially incorrect words or utterances and support downstream review, filtering, and data selection (Oneață et al., 2021). Confidence estimation can also degrade under acoustic or linguistic domain mismatch, motivating domain-aware evaluation and calibration (Li et al., 2022). Our setting differs from token- or word-level ASR confidence estimation because the monitored object is a continuously emitted extraction waveform and the decision is made from a fixed three-second causal prefix.

Responsible evaluation and cross-domain reliability. Responsible speech-system evaluation requires more than reporting average performance. Subgroup analyses must account for confounding factors, speaker heterogeneity, and incomplete demographic evidence to avoid unsupported fairness conclusions (Liu et al., 2022). We therefore audit intervention precision, failure coverage, false-intervention burden, target circularity, metadata coverage, and cross-domain consistency. Because the available demographic metadata are selective, we do not claim population-level fairness; instead, we report where the evidence supports reliability conclusions and where it remains insufficient.

A further distinction is required between within-cohort cross-validation and frozen external transport. A monitor can rank examples successfully within its development distribution while failing after changes in acoustic conditions, extraction architecture, or outcome definition. We therefore separate trial-disjoint internal evaluation from a later frozen-model study using independently computed system-specific reconstruction labels, and analyse the feature-distribution shift associated with any loss of transportability.

3 Causal Monitoring Framework

3.1 Causal Monitoring Setup

We use the beginning of a TSE output to estimate the risk that the completed output will be labelled adverse. For each trial, let x denote the mixture waveform, e the target-speaker enrollment utterance, and y the waveform emitted by a TSE system. The reduced monitor observes only an early prefix $y_{1:t}$ and estimates

$$r_t = f_\theta(\phi(y_{1:t})), \quad (1)$$

where $\phi(\cdot)$ extracts acoustic evidence available from the emitted waveform up to time t , and $r_t \in [0, 1]$ is the estimated failure risk. The enrollment utterance is used by the upstream TSE system to produce y , but is not an input to the reduced 63-feature monitor used in the principal analyses.

The causal monitor operates only on the waveform emitted by the evaluated TSE system. Here, “causal” means that it uses only audio available up to time t ; it does not imply causal inference. The monitor neither accesses future samples nor generates, selects, or replaces output waveforms. Its prediction is intended only for warning, abstention, escalation, or downstream review.

As illustrated in Figure 1, the emitted waveform is analysed using one-second causal windows updated every 0.25 seconds. These updates provide progressively more acoustic evidence and are not separate failure decisions. Our main evaluation uses the first 3 seconds of the output. This $t = 3$ s operating point was fixed before evaluation and is not claimed to be optimal. We additionally evaluate prefixes of 1, 2, 3, and 4 s to examine how early useful warning information appears and whether the results depend strongly on the 3-s choice.

Because independent failure labels are unavailable for all discovery outputs, we use the archived binary variable `proxy_adverse_outcome`. It combines seven complete-output indicators: silence, clipping, energy mismatch, spectral flatness, temporal instability, output–mixture inconsistency, and cross-system disagreement. The indicators are equally weighted after component scaling estimated on the predefined 7,812-output development split. An output is labelled proxy-adverse when the resulting score exceeds the fixed development-set 80th-percentile threshold of 0.7830. Neither the combination rule nor the threshold is selected using predictor performance. The proxy therefore represents potentially degraded or atypical TSE behaviour rather than an independently verified physical or perceptual failure label.

Some monitoring features may overlap with the indicators used to construct the proxy and thereby inflate predictive performance. To assess this target circularity, we compare three feature configurations. The *full acoustic set* contains all eligible causal variables; the *target-overlap-only set* contains variables directly overlapping with proxy construction; and the *reduced non-overlap set* excludes the target column, compos-

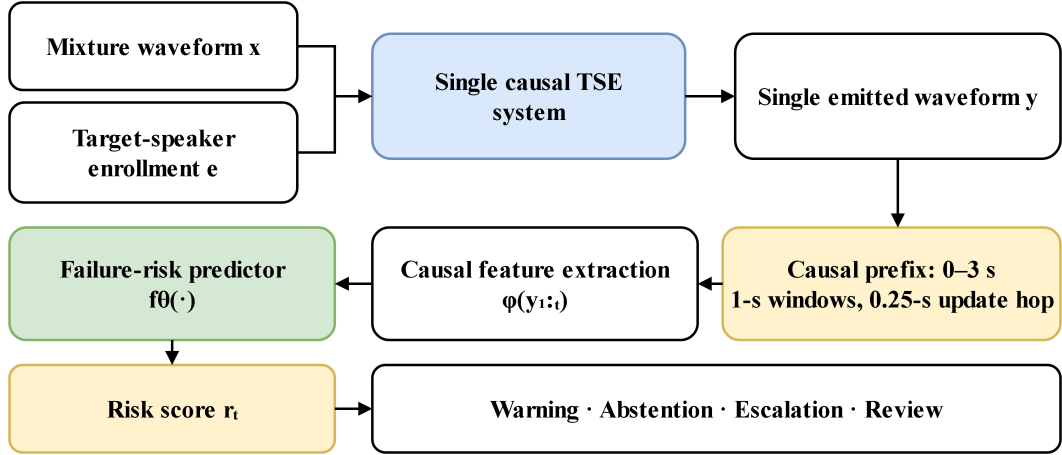


Figure 1: Single-stream causal monitoring framework. A fixed TSE system uses the mixture waveform and target-speaker enrollment utterance to produce one emitted waveform. The reduced monitor then uses only acoustic evidence available within the first three seconds of that waveform to estimate failure risk. It does not access future audio, generate alternative waveforms, select complete utterances, or replace the emitted output.

ite risk score, seven component scores, mixture-derived variables, and all other identified target-overlapping causal variables.

The reduced non-overlap set contains 63 features describing emitted-waveform amplitude, spectral-band ratios, spectral shape, zero-crossing behaviour, and temporal summaries over the causal prefix. It is “reduced” because target-overlapping variables are excluded, not because features were selected using evaluation-label performance. This configuration is used for all operating-policy analyses.

The predictor f_θ is a class-balanced L2-regularized logistic regression model with $C = 1.0$, fitted using the liblinear solver after median imputation, missing-value-indicator augmentation, and standard scaling. The same preprocessing and classifier are used across all three feature sets; only the input representation changes. Logistic regression provides a simple, interpretable basis for comparing the feature configurations.

3.2 Evaluation Protocol

We evaluate 11,946 outputs from 1,991 trials, where each trial is one mixture–target–enrollment instance evaluated by six TSE systems. The data span five conversational corpora: AISHELL-4, AMI, AliMeeting, CHiME6, and DipCo. The systems are USEF_SepFormer_WSJ0_2mix, USEF_TFGridNet_WSJ0_2mix, spk_emb_100, spk_emb_causal_100, tfmap_context_100, and tfmap_context_causal_100. Each trial contributes one output per system, yielding

$1,991 \times 6 = 11,946$ outputs.

Because the six outputs from a trial share the same mixture, target, and enrollment, they are kept in the same fold. We use five-fold trial-disjoint cross-validation repeated ten times and average the ten out-of-fold predictions for each output to obtain its final risk score. Let $z_i \in \{0, 1\}$ denote the proxy-defined discovery outcome.

Although cross-system disagreement contributes to offline target construction, the deployed monitor observes only the emitted waveform. The target, composite score, component scores, and all identified target-overlapping causal variables are excluded from the reduced non-overlap representation. We refer to outputs with $z_i = 1$ as proxy-defined failures; this does not imply independently verified perceptual or task failure. No independent measure of intelligibility, speaker preservation, timing, reconstruction quality, perceptual quality, or human listening quality is available for all 11,946 discovery outputs.

For frozen external evaluation, we use 500 trials and four TSE systems, yielding 2,000 system-output instances. Each has an independently computed system-specific scale-invariant signal-to-distortion ratio improvement (SI-SDRi), which measures whether extraction improves the target signal relative to the original mixture. External failure is defined as $\text{SI-SDRi} \leq 0$ dB, yielding 129 failures: 27 for spk_emb_100, 79 for spk_emb_causal_100, 9 for tfmap_context_100, and 14 for tfmap_context_causal_100. The saved re-

duced monitor is applied without refitting, feature reselection, calibration, or threshold adjustment. We also evaluate the same predictions against the original proxy-defined outcome on these 2,000 outputs, helping distinguish the effects of distribution shift and failure definition.

An intervention budget b denotes the fraction of outputs that can be reviewed or escalated. Outputs are ranked by predicted risk, and the highest-risk fraction b is selected. Internal analyses use averaged trial-disjoint out-of-fold scores, whereas external analysis uses scores from the unchanged frozen monitor. Let $\mathcal{S}_b$ denote the selected set and $\ell_i \in \{0, 1\}$ the relevant cohort-specific failure label: proxy-defined internally and SI-SDRi-defined externally. Intervention precision and failure coverage are

$$\begin{aligned} \text{Prec}(b) &= \frac{\sum_{i \in \mathcal{S}_b} \ell_i}{|\mathcal{S}_b|}, \\ \text{Cov}(b) &= \frac{\sum_{i \in \mathcal{S}_b} \ell_i}{\sum_i \ell_i}. \end{aligned} \quad (2)$$

Precision is the proportion of selected outputs that are failures, while failure coverage is the proportion of all failures captured within the budget. This budgeted-selection perspective is related to selective classification and reject-option methods (El-Yaniv and Wiener, 2010). Selective classification has also been studied for deep neural networks (Geifman and El-Yaniv, 2017).

We define coverage lift as $\text{Lift}(b) = \text{Cov}(b)/b$, since random selection has expected failure coverage b . Thus, a lift of $3.63\times$ at a 15% budget indicates 3.63 times the failure coverage expected under random selection. We define the false-intervention rate (FIR), equivalent to the false-positive rate when intervention is treated as the positive decision, as

$$\text{FIR}(b) = \frac{\sum_{i \in \mathcal{S}_b} (1 - \ell_i)}{\sum_i (1 - \ell_i)}. \quad (3)$$

The FIR is therefore the proportion of negative outputs selected for intervention.

We report representative budgets of 5%, 15%, and 30%; the cross-corpus and cross-system analysis uses the predefined set $\{15\%, 20\%, 30\%\}$. All ranking rules and budgets are fixed before subgroup comparisons and applied unchanged across corpora and TSE systems.

4 Results

4.1 Early Proxy-Adverse-Outcome Prediction

We first ask whether the first three seconds of the emitted TSE waveform contain sufficient information to predict whether the completed output will be labelled proxy-adverse. All predictions are out-of-fold and trial-disjoint, so outputs derived from the same underlying trial do not appear in both training and evaluation partitions. Of the 11,946 outputs, 2,181 are labelled proxy-adverse, corresponding to a prevalence of 18.26%. Because the expected AUPRC of a random ranking equals the positive-class prevalence, the random AUPRC baseline is 0.1826.

To assess the effect of feature–target overlap, all three feature configurations use the same class-balanced logistic-regression classifier, preprocessing pipeline, proxy target, repeated trial-disjoint folds, and evaluation procedure; only the input feature set changes. As shown in Table 1, the full acoustic feature set achieves an AUROC of 0.8991 and an AUPRC of 0.7613. Restricting the input to features identified as overlapping with proxy construction yields an AUROC of 0.8851 and an AUPRC of 0.7316, indicating that these features alone contain substantial information about the automatically constructed outcome. After removing all identified overlapping features, the reduced non-overlap model achieves an AUROC of 0.8062 and an AUPRC of 0.6606, compared with a random AUPRC baseline of 0.1826. Thus, substantial predictive signal remains after removal of the identified feature–target overlap.

Bootstrap comparisons show that the full model exceeds the reduced model by 0.0929 AUROC (95% CI: [0.0810, 0.1053]) and 0.1007 AUPRC (95% CI: [0.0862, 0.1163]). Both confidence intervals exclude zero, showing a clear contribution from the identified target-overlapping features while confirming that the reduced model retains useful ranking ability after their removal.

As a post-hoc sensitivity analysis, we separately evaluate full-prefix models using 1, 2, 3, and 4 seconds of observed audio. Their holdout AUROCs are 0.828, 0.832, 0.873, and 0.863, respectively, with corresponding AUPRCs of 0.358, 0.434, 0.498, and 0.495. These results indicate that predictive information is present before three seconds and that extending the observed prefix beyond three seconds does not improve performance

Feature set	AUROC	95% CI	AUPRC	95% CI
Full acoustic	0.8991	[0.8900, 0.9078]	0.7613	[0.7421, 0.7796]
Target-overlap only	0.8851	[0.8750, 0.8945]	0.7316	[0.7108, 0.7506]
Reduced non-overlap	0.8062	[0.7914, 0.8215]	0.6606	[0.6359, 0.6848]

Table 1: Trial-disjoint prediction performance at the predefined three-second causal decision point. Confidence intervals are obtained using 3,000 trial-cluster bootstrap repetitions. The reduced non-overlap model is used for all subsequent operating-policy analyses.

Budget	Precision	Coverage	Lift	FIR
5%	93.14%	25.54%	5.10×	0.42%
15%	66.29%	54.47%	3.63×	6.19%
30%	43.83%	72.03%	2.40×	20.61%

Table 2: Representative budget-constrained operating points for the reduced non-overlap model. FIR denotes the false-intervention rate.

in this sensitivity analysis. These models constitute a separate sensitivity experiment and should not be directly compared with the principal trial-disjoint reduced-model estimates in Table 1.

4.2 Budget-Constrained Utility

We next evaluate whether the reduced non-overlap model can prioritize proxy-adverse outputs when only a limited fraction of outputs can be reviewed or escalated. Outputs are ranked by their out-of-fold risk scores, and the highest-risk fraction is selected for intervention. The budget denotes the fraction of outputs that can be reviewed or escalated; for example, a 15% budget corresponds to selecting the highest-risk 15% of outputs. These budgets represent illustrative capacity constraints rather than deployment recommendations.

Table 2 reports three representative operating points. At a 5% budget, the monitor achieves 93.14% precision and captures 25.54% of proxy-defined failures, corresponding to a $5.10\times$ lift over random selection and a false-intervention rate of 0.42%. At a 15% budget, precision is 66.29% and failure coverage is 54.47%, corresponding to a $3.63\times$ lift. Increasing the budget to 30% raises failure coverage to 72.03%, while precision decreases to 43.83%, lift decreases to $2.40\times$, and the false-intervention rate increases to 20.61%.

4.3 Within-Cohort Cross-Corpus and Cross-System Analysis

We next apply the same predefined 15%, 20%, and 30% review policies across the five discovery corpora and six TSE systems, without corpus-

or system-specific retraining or threshold adjustment. The monitor captures more proxy-defined failures than expected under random selection in every corpus and system. At the 15% budget, the lowest corpus-level lift is $1.71\times$ on CHiME6, while the lowest system-level lift is $2.59\times$ for `spk_emb_100`.

Performance nevertheless varies across corpora and systems. At the 15% budget, AISHELL-4 achieves 64.71% precision and 45.83% failure coverage, whereas AliMeeting achieves 15.70% failure coverage. At the 20% budget, DipCo achieves 19.62% coverage, and at the 30% budget, AliMeeting precision is 17.07%. The comparatively large lift observed for AISHELL-4 should be interpreted in light of its low proxy-adverse prevalence (1.67%), and therefore jointly with precision and coverage. Overall, the same risk-ranking policy performs above random selection across all discovery corpora and systems, although its utility varies substantially across these groups.

4.4 Frozen Multi-System External Validation

We next apply the frozen reduced monitor to 2,000 unseen outputs from 500 trials and four TSE systems. External failure is independently defined as $\text{SI-SDR}_i \leq 0$ dB, yielding 129 failures and a prevalence of 6.45%. No refitting, feature reselection, calibration, threshold search, or hyperparameter adjustment is performed.

Against the independently defined SI-SDR_i outcome, the frozen monitor achieves a pooled AUROC of 0.5354 (95% CI: [0.4776, 0.5914]) and an AUPRC of 0.0700 (95% CI: [0.0552, 0.0977]), compared with a random AUPRC baseline of 0.0645. At a 15% review budget, the monitor selects 300 outputs and identifies 19 of the 129 SI-SDR_i-defined failures, corresponding to 6.33% precision, 14.73% failure coverage, and a $0.982\times$ lift over random selection. Per-system AUROC ranges from 0.5172 to 0.5494; because the system-specific failure counts are small, we treat these estimates as descriptive and base the external con-

Evaluation	Outcome	Outputs	Failures	AUROC	AUPRC	Random AUPRC
Internal OOF	Proxy-defined	11,946	2,181	0.8062 [0.7914, 0.8215]	0.6606 [0.6359, 0.6848]	0.1826
Frozen external	SI-SDRi ≤ 0 dB	2,000	129	0.5354 [0.4776, 0.5914]	0.0700 [0.0552, 0.0977]	0.0645

Table 3: Internal trial-disjoint and frozen external evaluation of the reduced causal monitor. The evaluations use different outcome definitions and therefore address different validity questions; their performance values should not be interpreted as a controlled comparison of the same prediction task.

clusion primarily on the pooled analysis. Overall, the frozen monitor does not reliably discriminate SI-SDRi-defined failures in this external cohort.

To examine whether this result reflects only a change in the input distribution, we also evaluate the unchanged predictions against the original proxy-defined outcome on the same 2,000 external outputs. Under this outcome definition, the monitor achieves an AUROC of 0.8092 and an AUPRC of 0.8481. Agreement between the proxy-defined and SI-SDRi-defined failure labels is weak ($\kappa = 0.028$, $MCC = 0.064$). Thus, the near-chance performance against SI-SDRi cannot be attributed to distribution shift alone; differences in the definition of failure are also associated with the observed loss of discrimination.

4.5 Distribution-Shift and Metadata Audits

A post-evaluation audit reveals substantial differences between the reduced monitoring features in the discovery and external cohorts. Of the 63 features, 24 have an absolute standardized mean difference of at least 0.50, 45 have a population stability index of at least 0.25, and 40 have a Kolmogorov–Smirnov statistic of at least 0.20. The median predicted risk increases from 0.3092 in the discovery cohort to 0.7448 in the external cohort. No individual feature achieves an oriented external AUROC of 0.65 or greater; the highest is 0.6070. These findings confirm substantial feature-distribution shift between the two cohorts. Together with the external proxy analysis above, they indicate that the external SI-SDRi result occurs in the presence of both feature shift and a change in outcome definition; the present analysis does not isolate their individual contributions.

We also assess the extent to which the available metadata support demographic analyses. Authentic language or gender metadata are available for only 355 of the 1,991 discovery trials, corresponding to 17.83% trial coverage. This subset is selective and cannot be assumed to represent the full discovery cohort. The available evidence therefore

does not support population-level claims about demographic fairness or disparity.

Finally, the discovery and external evaluations address different notions of adverse outcome. The discovery analysis uses the proxy-defined label because independent measures of intelligibility, target-speaker preservation, timing, reconstruction quality, perceptual quality, or human listening quality are unavailable for all 11,946 outputs. The external cohort provides independently computed SI-SDRi for 2,000 outputs from four systems, but SI-SDRi represents a single reconstruction-based definition of failure. The two evaluations should therefore be interpreted as complementary tests of within-setting prediction and transport to an independently defined outcome, rather than as evaluations of an identical prediction task.

5 Discussion

The results separate two questions that can otherwise be conflated. First, early causal acoustic evidence contains useful predictive information within the proxy-defined discovery setting. After removing features identified as overlapping with proxy construction, the reduced model retains an AUROC of 0.8062 and an AUPRC of 0.6606. The monitor also provides useful budget-constrained ranking within this cohort, including a $3.63\times$ lift over random selection at the representative 15% review budget.

Second, performance within the proxy-defined discovery setting does not establish reliability for independently defined failure outcomes on unseen data. Although the same risk-ranking policy performs above random selection across all five discovery corpora and six discovery systems, the frozen monitor performs near chance against SI-SDRi-defined failure on the external four-system cohort (AUROC 0.5354; AUPRC 0.0700). The external cohort exhibits substantial feature-distribution shift. However, on the same external outputs, the unchanged monitor remains strongly predictive of the original proxy-defined outcome

(AUROC 0.8092; AUPRC 0.8481), while agreement between the proxy and SI-SDRi failure labels is weak. Together, these findings indicate that the near-chance SI-SDRi performance cannot be attributed to distribution shift alone; the change in failure definition is also associated with the loss of discrimination. The present analysis does not disentangle the individual contributions of these two factors.

Calibration may alter score interpretation and operating thresholds, but it cannot recover ranking information when discrimination against the target outcome is near chance. Reliable transport may therefore require more robust representations, independently grounded training targets, or adaptation to the intended deployment setting. These possibilities remain to be evaluated.

Overall, causal early warning provides a useful mechanism for prioritizing proxy-adverse outputs within the setting in which the monitor is developed, but reliable monitoring across new systems, data distributions, and independently defined failure outcomes remains unresolved. The monitor should therefore be treated as an auxiliary prioritization and review tool rather than as a deployment-ready failure detector.

6 Limitations

First, the discovery outcome is defined automatically using seven equal-weight proxy diagnostics rather than independent perceptual or task-level labels. This proxy is not assumed to be an optimal definition of failure, and alternative weightings or outcome definitions may produce different results.

Second, the external evaluation is limited to 500 trials from four systems and uses SI-SDRi as a single reconstruction-oriented failure criterion. SI-SDRi does not capture other important aspects of TSE performance, such as intelligibility, target-speaker preservation, timing, perceptual quality, or human preference. Moreover, because the internal and external evaluations use different failure definitions, their performance values are not directly comparable. The external cohort also exhibits substantial feature shift. Although evaluation against the original proxy suggests that both data shift and the change in failure definition are relevant, the present study does not disentangle their individual effects. We also do not establish whether more complex nonlinear models would improve external transport.

Third, authentic demographic metadata are available for only 355 of the 1,991 discovery trials. The resulting analysis is therefore a limited demographic audit and cannot establish population-level fairness or disparity.

Finally, the study evaluates early warning and review prioritization, not the effectiveness of subsequent interventions. It does not establish that acting on an alert prevents failure, improves the emitted waveform or standard TSE metrics, or leads to better outcomes for human listeners.

7 Ethics and Responsible Use

A failure-risk monitor may reduce the need to inspect every output, but it also introduces operational trade-offs. False alerts can lead to unnecessary review or abstention, whereas missed alerts may allow unreliable outputs to pass without intervention. The operating point should therefore be selected according to the relative costs of these errors and the review capacity available in the intended deployment setting.

The monitor should not be interpreted as a fairness certificate, a substitute for human evaluation, or evidence of safety in high-consequence settings. Its use should be supported by independent outcome labels, representative data, and domain-specific cost assumptions. In particular, the frozen external results show that performance can degrade substantially under changes in the data distribution and failure definition, and that recalibration alone should not be assumed to restore reliable performance. Deployment should therefore be preceded by independently labelled, prospective validation on the intended systems, populations, and recording conditions.

8 Conclusion

We presented a single-stream causal early-warning framework that uses only the first three seconds of a TSE system’s emitted waveform. Under trial-disjoint evaluation against the proxy-defined discovery outcome, the reduced non-overlap monitor achieved an AUROC of 0.8062 and an AUPRC of 0.6606. At a 15% intervention budget, it identified 54.47% of proxy-defined failures with 66.29% precision, corresponding to a $3.63\times$ lift over random prioritization.

The unchanged monitor did not reliably discriminate independently defined SI-SDRi failures on 2,000 external outputs from four systems. Under

this evaluation, AUROC was 0.5354 and AUPRC was 0.0700, compared with a random AUPRC baseline of 0.0645. The external cohort also exhibited substantial feature-distribution shift. Importantly, on the same external outputs, the unchanged monitor retained strong discrimination for the original proxy-defined outcome (AUROC 0.8092; AUPRC 0.8481). Together, these findings indicate that the near-chance SI-SDRi performance cannot be attributed to distribution shift alone; the change in failure definition is also associated with the loss of discrimination. The present analysis does not disentangle the individual contributions of these factors.

Overall, our findings support causal early warning as a useful within-cohort prioritization mechanism, while highlighting the limitations of treating such a monitor as a universal, deployment-ready failure detector. Future work should focus on learning shift-robust causal representations and on prospective validation against independently defined outcomes that capture perceptual quality, speaker preservation, timing, and intelligibility.

Generative AI Assistance

Generative language tools assisted language editing and organization. The authors verified all experiments, analyses, references, and final text.

References

- Marc Delcroix, Shinji Watanabe, Tsubasa Ochiai, Keisuke Kinoshita, Shigeki Karita, Atsunori Ogawa, and Tomohiro Nakatani. 2019. [End-to-end Speaker-Beam for single channel target speech recognition](#). In *Proceedings of Interspeech*, pages 451–455.
- Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani. 2018. [Single channel target speaker extraction and recognition with SpeakerBeam](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5554–5558.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuyne. 2020. [ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN-based speaker verification](#). In *Proceedings of Interspeech*, pages 3830–3834.
- Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641.
- Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30.
- Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159.
- Qiujia Li, Yu Zhang, David Qiu, Yanzhang He, Lianliang Cao, and Philip C. Woodland. 2022. [Improving confidence estimation on out-of-domain data for end-to-end speech recognition](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6537–6541.
- Kai Liu, Ziqing Du, Xucheng Wan, and Huan Zhou. 2023. [X-SepFormer: End-to-end speaker extraction network with explicit optimization on speaker confusion](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Zhe Liu, Irina-Elena Veliche, and Fuchun Peng. 2022. [Model-based approach for measuring the fairness in ASR](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6532–6536.
- Yi Luo and Jianwei Yu. 2023. [Music source separation with band-split RNN](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901.
- Dan Oneață, Alexandru Caranica, Adriana Stan, and Horia Cucu. 2021. [An evaluation of word-level confidence estimation for end-to-end automatic speech recognition](#). In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 258–265.
- Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023. [WeSpeaker: A research and production oriented speaker embedding learning toolkit](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Ke Zhang, Junjie Li, Shuai Wang, Yangjie Wei, Yi Wang, Yannan Wang, and Haizhou Li. 2025. [Multi-level speaker representation for target speaker extraction](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Zifeng Zhao, Dongchao Yang, Rongzhi Gu, Haoran Zhang, and Yuexian Zou. 2022. [Target confusion in end-to-end speaker extraction: Analysis and approaches](#). In *Proceedings of Interspeech*, pages 5333–5337.