

Prosodic Vulnerability in Remote Simultaneous Interpreting: Framework and Staged Research Agenda

Pio Savelli

Jilin International Studies University, Changchun, Jilin, 130117, China

Abstract: Remote simultaneous interpreting (RSI) depends on transmission chains that may modify temporal, F_0 -related, and intensity/level properties of speech. Yet global quality judgements and platform labels cannot explain why the same channel condition affects signals, listeners, and tasks differently. This conceptual article defines prosodic vulnerability as the estimated susceptibility of a functionally relevant prosodic contrast, in a specified source signal and for a specified listener population and task, to a reduction in functional cue availability under a specified channel transformation relative to a prespecified baseline. The observed baseline-to-transformation contrast in functional cue availability estimates this susceptibility; acoustic modification remains a distinct antecedent variable, whereas effort and task performance are distinct downstream outcomes. The framework links source-signal organization, channel transformation, acoustic cue modification, functional cue availability, compensatory processing, and observable outcomes. It predicts source-by-channel differences, greater vulnerability when perceptual salience or redundancy is low, and possible accuracy–effort dissociations when compensation competes with simultaneous-interpreting demands. A staged research agenda first validates the signal-internal pathway in controlled perception, then examines source-to-interpreter and interpreter-to-audience transmission separately and, where feasible, factorially. For controlled tests of the signal-internal pathway, the framework requires documented transformations, matched input–output signals, independent functional-cue measures, prespecified outcomes, and explicit falsification criteria; field monitoring without a clean reference supports only provisional technical observations, not an inference of prosodic vulnerability. It provides a conditional basis for assessing platform configurations, calibrated delivery adaptation, transmitted vocal delivery, and prosody-aware tools without treating speakers or accents as intrinsically vulnerable or transferring responsibility for inadequate technical conditions to interpreters.

Keywords: Remote simultaneous interpreting; Prosodic vulnerability; Digitally mediated speech; Speech signal degradation; Functional cue availability; Speech intelligibility; Listening effort

DOI: 10.62639/sspjiss18.20260308

1. Introduction

Remote simultaneous interpreting (RSI) is not simply conventional simultaneous interpreting performed at a distance. It places the interpreter within a technologically mediated communicative environment in which access to source speech depends on a transmission chain that may include microphone capture, automatic audio processing, encoding, network transmission, decoding, and playback (Mouzourakis, 2006). Each stage can modify temporal, spectral, F_0 -related, and intensity/level properties of the signal. At the same time, RSI cannot be reduced to audio transmission alone: visual access, interactional coordination, workspace ergonomics, and the cognitive demands of simultaneous interpreting also shape performance (Cheung, 2024; Hale et al., 2022; Moser-Mercer, 2005; Salaets & Brône, 2023). The present article therefore isolates the acoustic-perceptual dimension of RSI while treating it as one component of a broader communicative ecology.

(Manuscript NO.: JISS-26-8-F001)

About the Author

Pio Savelli (2002–), male, Italian, Lecturer, Master's degree.

Global transmission quality is commonly assessed through standardized subjective procedures such as those specified in ITU-T Recommendation P.800 (International Telecommunication Union [ITU], 1996; Raake, 2006). Such a view is insufficient for explaining why the same transmission condition may have different perceptual consequences for different speech signals. Digital processing does not operate on abstract linguistic content, but on acoustic signals whose cues differ in salience, redundancy, distribution, and functional importance (Aylett & Turk, 2004; Holt & Lotto, 2006). A transformation that is perceptually inconsequential for one signal may reduce the accessibility of information in another. The relevant analytical problem is therefore not simply whether a signal has been degraded, but how source-signal organization interacts with a specified channel transformation, which acoustic cues are modified, whether functional cue availability is reduced, what compensatory processing is recruited, and which observable outcomes follow.

Prosody is especially relevant to this problem because timing, pausing, prominence, intensity relations, and fundamental-frequency movement contribute to speech segmentation, information structure, anticipation, and pragmatic interpretation (Cutler et al., 1997; Wagner & Watson, 2010). Under favourable listening conditions, these resources operate together with segmental, lexical, syntactic, and contextual information. When some prosodic cues become less salient, listeners may compensate by relying more heavily on other sources of information. Successful word recognition may therefore be maintained at the cost of reduced comprehensibility or increased listening effort. Intelligibility, perceived ease of understanding, processing effort, and interpreting performance must consequently be treated as related but distinct outcomes (Derwing & Munro, 2015; McGarrigle et al., 2014; Pichora-Fuller et al., 2016).

These distinctions are particularly important in RSI. Any additional effort devoted to recovering or reconstructing the source message may compete with the concurrent demands of comprehension, retention, reformulation, and target-speech production (Gile, 2009; Seeber, 2011). It is therefore plausible that a reduction in the functional availability of prosodic cues could have greater consequences in an interpreting task than in ordinary listening. This possibility, however, should be treated as an empirical prediction rather than as an established effect. Evidence that an acoustic parameter changes during transmission would not, by itself, demonstrate a meaningful loss of prosodic information, an increase in cognitive load, or a decline in interpreting performance.

To address this problem, the article defines prosodic vulnerability as the estimated susceptibility of a functionally relevant prosodic contrast, in a specified source signal and for a specified listener population and task, to a reduction in functional cue availability under a specified channel transformation relative to a prespecified baseline. The construct is relational, graded, cue-specific, channel-specific, listener-dependent, and task-dependent. It denotes susceptibility to cue-availability reduction; realized acoustic modification, listener effort, and task performance are separate variables used to test the proposed pathway. It must therefore not be used to classify speakers, accents, or voices as intrinsically “vulnerable” or “robust.”

To our knowledge, existing research on codec effects on emotional prosody (Niebuhr & Siegert, 2023), prosodic processing in simultaneous interpreting (Du & Wang, 2025), and RSI configurations and interpreter outcomes (Chmiel et al., 2025) does not provide a single framework requiring independent specification of the source contrast, channel transformation, acoustic modification, functional-cue measure, and downstream RSI outcome. The novelty claimed here is therefore methodological and relational, not that these constituent phenomena have been ignored.

The article develops a testable conceptual framework rather than reporting new findings. It defines the construct, specifies the proposed channel–cue–function–outcome relations, and presents a staged research agenda for controlled perception and RSI. It offers an analytical vocabulary and reporting requirements, not a validated mechanism or a fully specified experimental protocol.

2. Conceptual Foundations

(1) Prosody as a functional resource for speech processing

Prosody refers here to temporal organization, F_0 -related structure, and intensity/level patterning in speech. Pause placement, duration, rhythm, F_0 movement, and relative intensity help listeners locate boundaries, identify prominence, relate successive units, and interpret syntactic or pragmatic structure (Chakhachiro, 2016; Cutler et al., 1997; Wagner & Watson, 2010). Acoustic intensity or level, perceived loudness, and dynamic-range processing are distinct: they concern, respectively, a signal measure, a perceptual attribute, and a transformation applied to the signal. Prosodic functions are conveyed by overlapping cues whose importance varies across utterances, speakers, languages, listeners, and tasks.

This functional perspective is particularly relevant to interpreting, where the listener must identify message structure while processing incoming speech, retaining prior material, and formulating the target text. Prosodic grouping and prominence may make relations within the unfolding message more accessible (Ahrens, 2005). Experimental work in simultaneous interpreting with text likewise indicates that controlled prosodic features can affect trainee interpreters' cognitive effort, auditory–visual coordination, and interpreting output, with some features facilitating prediction and coordination and others hindering it (Du & Wang, 2025). This evidence establishes the task relevance of prosodic information but, because digital transmission was not manipulated, does not demonstrate channel-induced cue loss. The functional contribution of such cues, however, cannot be inferred from acoustic form alone (Lenglet & Michaux, 2020). A measurable pitch movement or intensity contrast is not necessarily perceptually usable; conversely, a change in one parameter need not cause information loss if other cues remain salient or redundant.

Functional cue availability is the degree to which a targeted prosodic contrast remains perceptually usable for a specified listener population and task under specified signal conditions. It is estimated from performance on a prespecified cue-diagnostic task; this measure must remain distinct from the downstream outcome it is expected to predict. Here, the task that conditions the functional-cue-availability estimate is the cue-diagnostic task; RSI is treated as a downstream application task in which the consequences of a validated cue-availability reduction are assessed separately. For each function, the study should specify the target contrast, cue configuration, task, scoring rule, baseline, reliability, and criterion for meaningful reduction after transformation. The same acoustic change may be inconsequential when redundant cues support the contrast but consequential when the function depends on a weak cue (Aylett & Turk, 2004; Holt & Lotto, 2006). Relative vulnerability must therefore be established under specified conditions.

(2) Digital signal mediation and functional cue availability

In RSI, the source signal reaches the interpreter through a chain that may include microphone capture, automatic gain control, noise suppression, echo cancellation, encoding, network transmission, decoding, and playback. ISO 24019:2022 addresses simultaneous-interpreting delivery platforms, while ISO 20109:2025 covers interpreting equipment and the quality of sound and image transmitted to interpreters and audiences (International Organization for Standardization [ISO], 2022, 2025). These standards set technical requirements but do not identify which linguistic cues a particular implementation alters or demonstrate a prosodic effect.

Bandwidth restriction, codec processing, packet loss, jitter management, buffering, and latency are not interchangeable forms of “digital degradation.” The model distinguishes a signal-internal pathway, comprising transformations that alter properties within an utterance and can be assessed with matched input–output signals, from an interactional pathway, comprising delay and coordination disruptions that may affect turn-taking or performance without changing internal prosodic organization. Latency belongs to the prosodic pathway only when it demonstrably alters within-utterance timing. Packet loss and buffering should likewise

be described by their observed effects rather than their technical labels alone (Mouzourakis, 2006; Raake, 2006).

Because commercial platforms combine adaptive codecs, user settings, network conditions, capture devices, and playback equipment, a platform name cannot serve as a reproducible experimental condition (Salaets & Brône, 2023). Claims about the signal-internal pathway therefore require documented technical parameters and matched input-output signals derived from the same baseline recording. In field conditions where a clean input reference cannot be obtained, the evidence is limited to descriptive or diagnostic monitoring and cannot by itself support an inference of prosodic vulnerability; the corresponding reporting requirements are specified in Section 4.2. Direct empirical evidence that codec processing can modify prosodic patterning is provided by Niebuhr and Siegert (2023). Using the same emotional utterances in an uncompressed reference condition and after processing with AMR-WB, MP3, Opus, and Speex, they found codec- and parameter-specific changes across temporal, F_0 -related, intensity, and spectral measures; non-tonal dimensions were generally more affected, and the acoustic pattern partly paralleled perceptual emotion-merging effects reported in their earlier work. This study is therefore a close empirical antecedent of the signal-internal pathway proposed here. It does not, however, test the complete evidential chain within a single design: functional cue availability, compensatory processing, and a distinct RSI outcome were not independently estimated.

The issue is consequently not whether digital mediation “damages prosody” in general, but whether a specified transformation reduces the functional availability of particular cues and predicts a relevant listener outcome. This separates four levels often conflated: technical alteration of the signal, measurable change in an acoustic cue, independently measured functional cue availability, and a distinct downstream perceptual or task-related outcome. Only a documented link across these levels can support an inference about prosodic vulnerability.

(3) Degraded listening, compensation, and effort

Speech comprehension under adverse conditions depends on the interaction between bottom-up evidence and knowledge-driven prediction. When acoustic information is incomplete or less reliable, listeners may rely more heavily on lexical probability, syntactic structure, discourse context, and prior experience. Such compensation can preserve recognition while increasing the attention or effort required. Accuracy alone is therefore an incomplete index of perceptual cost: similar recognition scores may conceal different listener demands (Mattys et al., 2012; Pelle, 2018; Rönnberg et al., 2013; Wild et al., 2012).

Intelligibility concerns what is recognized; comprehensibility, the perceived ease of understanding; listening effort, the resources mobilized; and interpreting performance, the accuracy and adequacy of the target text (Derwing & Munro, 2015; Kang et al., 2025; McGarrigle et al., 2014; Pichora-Fuller et al., 2016). These outcomes are related but not interchangeable. A listener may reconstruct a message accurately while reporting greater difficulty, and increased effort need not immediately reduce intelligibility. Research on adverse listening conditions supports treating intelligibility and effort as separable dimensions rather than alternative labels for one effect (Ershaid et al., 2024).

In RSI, compensatory listening may be especially consequential because source-speech processing occurs alongside retention, reformulation, monitoring, and target-speech production. If recovery of the incoming message requires additional resources, fewer may remain for these concurrent operations. Remote-interpreting research links mediated conditions with altered task demands, stress, cognitive load, and performance (Cheung, 2024; Chmiel et al., 2025; Moser-Mercer, 2005; Roziner & Shlesinger, 2010; Zhu & Aryadoust, 2022). Chmiel et al. (2025), for example, reported the lowest objective accuracy and self-assessed performance in the chat-based, non-co-located boothmate condition, with little difference between the co-

located and virtual-booth conditions; cognitive load and anxiety did not vary across conditions. Because their manipulation concerned RSI configuration rather than a documented acoustic transformation, those findings do not establish that prosodic attenuation caused the observed performance pattern. The claim advanced here is therefore conditional: reduced functional availability of prosodic cues may increase effort and affect interpreting performance when compensation competes with other time-sensitive processes.

Prosodic vulnerability intersects with, but is not reducible to, adverse-listening accounts (Mattys et al., 2012), ELU and FUEL (Pichora-Fuller et al., 2016; Rönnberg et al., 2013), cue-weighting and redundancy accounts (Aylett & Turk, 2004; Holt & Lotto, 2006), or interpreting models (Gile, 2009; Seeber, 2011). Its added requirement is to demonstrate a specified channel–cue–function relation independently before effort or RSI effects are attributed to prosodic degradation. Aylett and Turk (2004) and Holt and Lotto (2006) provide theoretical analogies; neither directly tests digitally transmitted prosody or RSI. Direct transmission evidence must be supplied for each technical transformation. The construct adds value only if it changes measurement or prediction relative to these adjacent accounts.

3. A Model of Prosodic Vulnerability

(1) Defining the construct

Prosodic vulnerability is an *ex ante* susceptibility construct. Realized acoustic modification, measured reduction in functional cue availability, and downstream listener or interpreting outcomes are distinct variables used to test the proposed pathway. The construct therefore concerns neither acoustic alteration alone nor a general impression that a transmitted voice sounds less clear.

Prosodic vulnerability is relational rather than intrinsic. It is not a permanent property of a speaker, accent, language variety, or voice type. The same source signal may show a negligible reduction in functional cue availability under one form of processing and a larger reduction under another, depending on which cues are modified and which functions they perform. Describing a speaker as globally “vulnerable” or “robust” would consequently obscure the interaction that the construct is intended to capture.

The construct is also graded and cue-specific: under a specified channel transformation, listener population, and task, the expected reduction in functional cue availability may be larger for temporal boundary cues than for tonal or intensity cues, or larger for prominence cues than for pause-related cues. Prosodic vulnerability should therefore be operationalized as an estimand: the expected reduction in an independently specified measure of functional cue availability under a defined channel transformation relative to a prespecified baseline, conditional on the source signal, listener population, and task. The observed cue-availability contrast estimates this quantity, whereas acoustic cue modification and listener responses remain, respectively, antecedent evidence and downstream outcomes rather than components of the construct.

Finally, the functional-cue-availability estimate depends on the listener population and on the cue-diagnostic task used to obtain it, whereas the downstream consequences of a validated reduction may vary across application tasks. The functional value of a cue may vary according to language proficiency, familiarity with the speaker or accent, contextual predictability, listening conditions, and the demands of the task. A reduction that has little observable effect in ordinary comprehension may become consequential when the listener must simultaneously retain, reformulate, and produce information, as in simultaneous interpreting.

(2) Components of the model

The proposed model connects six analytical levels: source-signal organization (Aylett & Turk, 2004; Holt

& Lotto, 2006; Wagner & Watson, 2010), channel transformation, acoustic cue modification, functional cue availability, compensatory processing (Pichora-Fuller et al., 2016; Rönnberg et al., 2013), and observable outcome (McGarrigle et al., 2014; Ward et al., 2017). These levels describe a hypothesized sequence rather than an already established causal mechanism.

The first level, source-signal organization, concerns the configuration of cues before digital mediation. Acoustic salience is the magnitude or distinctiveness of a cue in the signal; perceptual salience is listener- and task-dependent and requires independent behavioural evidence; informational redundancy is the extent to which multiple cues or contextual sources support the same distinction. These variables are not interchangeable. Resilience may arise when several independently validated temporal, F_0 -related, or intensity cues converge.

The second level, channel transformation, comprises the documented technical processes applied to the signal. These may involve spectral restriction, codec processing, dynamic-range modification, packet loss, packet-loss concealment, noise suppression, or other temporally and acoustically relevant operations. Platform names cannot substitute for this level of description. In controlled tests of the signal-internal pathway, the transformation must be identified through reproducible settings and matched input-output measurements; field observations without a clean reference may document technical events or unusual signal behaviour, but they cannot establish acoustic cue modification or the full mechanism.

The third level, acoustic cue modification, refers to the measurable changes produced in the acoustic correlates of speech as the signal passes through the transmission channel. Depending on the type and severity of the transformation, these changes may include attenuation, temporal smearing or truncation, discontinuities in fundamental-frequency contours, altered intensity relations, reduced spectral detail, or disruption of fine-grained durational contrasts (Mattys et al., 2012; Raake, 2006). It should therefore be established through direct comparison of input and output signals rather than inferred from nominal technical settings. Its functional significance must be tested separately, because redundancy, contextual information, and cue reweighting may preserve usability despite measurable acoustic change.

The fourth level, functional cue availability, is a hypothesized mediator estimated from performance on a prespecified cue-diagnostic task that depends on the targeted cue. The resulting measure must remain distinct from the downstream outcome. The study must define the target contrast, task, scoring metric, baseline, reliability, and smallest meaningful reduction. Neither acoustic change nor downstream performance alone can measure this mediator.

The fifth level, compensatory processing, concerns the listener's response to reduced cue availability. Listeners may draw more extensively on lexical knowledge, syntactic expectations, discourse context, visual information, and predictive inference. Compensation may preserve recognition, but it may also increase the resources required to achieve it. Listener characteristics and contextual support therefore act as moderators rather than as peripheral sources of variation.

The sixth level comprises observable outcomes. These include transcription-based intelligibility, perceived comprehensibility, listening effort, and interpreting performance. Because the four outcomes represent different dimensions of processing, they should be measured separately. In particular, stable intelligibility cannot be taken as evidence that degradation has had no effect if equivalent performance is maintained through greater effort.

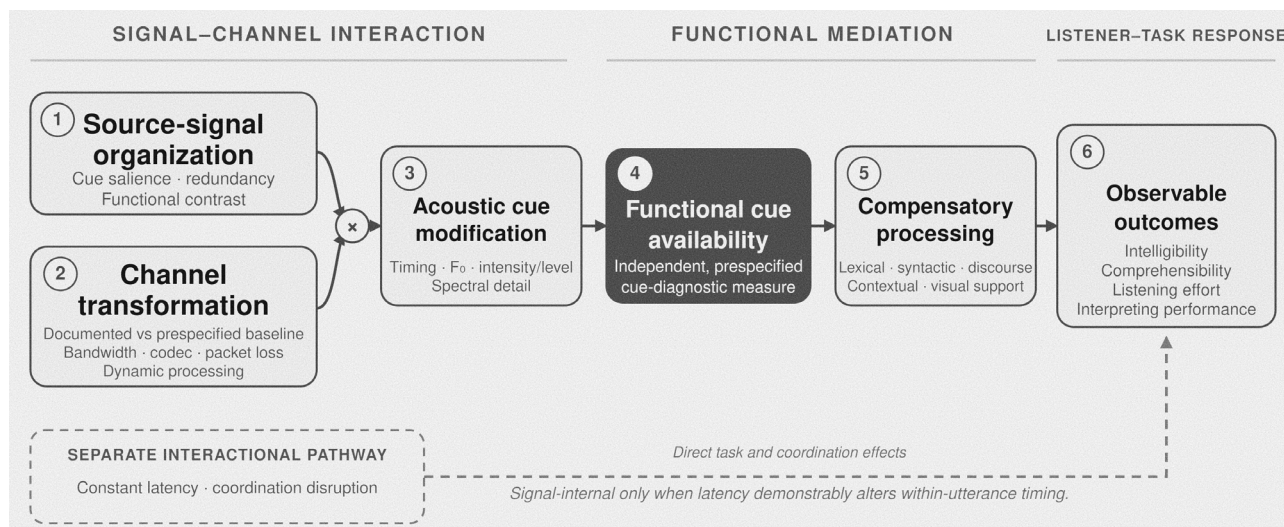


Figure 1. Conceptual model of prosodic vulnerability in digitally mediated speech. Source-signal organization interacts with a specified channel transformation to produce measurable acoustic cue modification. Downstream consequences depend on functional cue availability and compensatory processing before being reflected in intelligibility, comprehensibility, listening effort, or interpreting performance. The dashed path represents a separate interactional pathway through which constant latency or coordination disruption may affect observable outcomes without acoustic cue modification, unless within-utterance timing is demonstrably altered. Arrows represent hypothesized directional relations requiring separate empirical validation

(3) The differential vulnerability principle

The model is organized around a differential vulnerability principle: the perceptual consequences of a channel transformation depend on its interaction with the functional organization of the source signal. Digital mediation should therefore not be expected to produce a uniform reduction in speech quality across signals, cues, listeners, and tasks. This principle generates five testable propositions.

Proposition 1: Source-by-channel interaction. For a prespecified transformation and functional contrast, the baseline-to-transformation change in the functional-cue measure should vary with a prespecified continuous property of source-signal organization according to a preregistered directional prediction and functional form (e.g., linear, threshold, or nonlinear). Support requires the estimated interaction to meet the preregistered criteria for direction, functional form, and minimum magnitude; a statistically non-zero interaction alone is insufficient. The proposition is not supported if the interaction is negligible within the preregistered equivalence region, opposite in direction, or materially inconsistent with the specified form.

Proposition 2: Saliency and redundancy. Lower independently measured perceptual saliency or informational redundancy should predict a larger reduction in the functional-cue measure. The proposition is not supported if the prespecified association is negligible or opposite in direction.

Proposition 3: Functional mediation. A transformation should affect a downstream outcome through reduced availability of the targeted cue. Acoustic change without a reduction in the independent functional-cue measure does not support this proposition; segmental, global-quality, attentional, and interactional alternatives must be assessed separately.

Proposition 4: Accuracy-effort dissociation. A transformation may increase effort beyond a prespecified smallest effect of interest while intelligibility remains within a non-inferiority bound. This supports a cost-performance dissociation, but compensation only if its independent criteria are also met.

Proposition 5: Task-dependent amplification. A validated cue-availability reduction should have a larger effect in RSI than in a matched perception task on the same prespecified outcome measured with the same scoring rule.

If identical measurement is not feasible, the comparison must use a preregistered common measurement model that links task-specific indicators to the same construct and establishes their cross-task comparability; standardizing effects derived from conceptually different outcomes is insufficient. Support requires a task-by-transformation interaction estimated on that shared outcome or common measurement scale. The proposition is not supported if the interaction is negligible or opposite in direction.

Equivalence regions, smallest effects of interest, directional predictions, functional forms, outcome definitions and, where needed, cross-task measurement models must be justified and preregistered for each study; they are not universal properties of the framework.

(4) Evidential requirements

A defensible claim of prosodic vulnerability requires evidence across several distinct layers. No individual layer is sufficient on its own.

Table 1. Evidential layers required for an inference of prosodic vulnerability

Evidential layer	Central question	Required evidence
Source-signal organization	Which cues convey the relevant function before mediation?	Prespecified acoustic measures and, where appropriate, baseline perceptual validation
Channel transformation	What happened technically to the signal?	Documented processing parameters, matched input–output signals, and repeated transmissions where conditions are adaptive
Acoustic cue modification	Did the relevant temporal, tonal, or dynamic properties change?	Comparative analysis of matched baseline and transformed stimuli
Functional cue availability	Did the change reduce the perceptual usefulness of the cue?	Cue-identification, boundary-detection, prominence, or comparable functional measures
Compensatory processing	Did reduced cue availability lead listeners to rely more heavily on contextual, predictive, or visual information?	Interactions between cue availability and prespecified manipulations of lexical, syntactic, discourse, or visual support, with converging behavioural or physiological measures where appropriate
Observable outcome	Was the change associated with a relevant consequence?	Separate measures of intelligibility, comprehensibility, listening effort, or interpreting performance

The strongest inference arises when a prespecified property of source-signal organization interacts with a randomized, documented channel transformation, evaluated using matched input–output signals, to produce measurable acoustic cue modification. Evidence should then show, in the prespecified temporal order, a reduction in functional cue availability estimated with an independent prespecified measure, independently assessed compensatory processing where that mechanism is hypothesized, and a prespecified change in a distinct downstream outcome. Causal mediation should be claimed only when mediator timing, confounding assumptions, model specification, and sensitivity analyses are explicit.

The model thus treats prosodic vulnerability as an empirically testable interaction rather than as an assumed characteristic of digitally mediated speech. The next section translates these requirements into a staged research agenda designed to distinguish channel effects, cue modification, compensation, and task performance.

4. Staged Research Agenda for Empirical Evaluation

The following staged research agenda translates the six analytical levels in Section 3 into controlled perception and RSI studies. Each level must be measured separately so that a baseline–transformed contrast is not mistaken for evidence of the mechanism.

(1) Source-signal sampling

An empirical test should sample speakers, linguistic items, listeners or interpreters, and, under adaptive

transmission, independent transmission runs. The design must include enough observations at each level to distinguish systematic source-by-channel interactions from speaker- or item-specific variation and must state the population to which each inference is intended to generalize. Source properties should remain continuous; differences between two speakers cannot justify “vulnerable” and “robust” categories.

The source recordings should contain materials in which the relevant prosodic functions can be identified independently of the channel condition. These may include boundary placement, prominence assignment, tonal grouping, and pragmatic distinctions. Items should be balanced for lexical frequency, syntactic complexity, length, segmental composition, and contextual predictability wherever these variables could affect recognition or compensation.

Read speech, prepared delivery, spontaneous speech, and interpreted output should be treated as distinct production modes. Interpreted output in particular has been shown to exhibit distinctive prosodic characteristics whose functional consequences cannot be assumed from acoustic form alone (Ahrens, 2005; Lenglet & Michaux, 2020). They may be included within the same research programme, but should not be pooled without explicitly modelling their differences. Read materials provide greater experimental control, whereas spontaneous or interpreted speech offers greater ecological validity. A staged design may therefore begin with controlled utterances and subsequently examine longer, more naturalistic discourse.

All baseline recordings should be made in a documented acoustic environment using specified microphones, interfaces, sampling rates, bit depths, microphone distances, and input levels. The unprocessed recordings should be retained in a lossless format. Selection of speakers, utterances, and prosodic targets must occur before listener outcomes are examined. This prevents items from being selected because they happen to display the predicted degradation effect.

(2) Channel manipulation

Initial experiments should isolate one primary channel transformation whose acoustic and perceptual effects can be evaluated independently. Possible transformations include bandwidth restriction, a specified codec and bitrate, dynamic processing, or controlled packet loss with a documented concealment procedure. If several transformations are examined, they should be introduced through a factorial or otherwise explicitly structured design rather than combined into an undifferentiated degradation condition.

Each transformed stimulus should be derived from the same lossless baseline signal. Codec, bitrate, sampling rate, transmitted bandwidth, packet-loss rate, jitter conditions, buffer settings, noise suppression, automatic gain control, and other active processing features should be reported whenever applicable. The actual output signal must be captured and analysed; intended settings alone do not demonstrate what transformation occurred.

Deterministic offline processing and adaptive real-time transmission should be distinguished (Raake, 2006). Offline processing permits exact replication of a transformation, whereas platform-mediated transmission may vary across runs because of network adaptation, processing updates, and interactions among devices and settings. Adaptive conditions consequently require multiple independent transmissions of every relevant condition, together with network and endpoint documentation.

A commercial platform name should not be treated as a stable experimental variable. If a platform is included, it must be operationalized through its version, date of testing, audio settings, processing features, user devices, network parameters, capture method, and repeated transmission runs. The purpose of such a condition is to examine a documented implementation, not to infer a permanent quality hierarchy among commercial products.

Constant latency should be tested separately in interactive or interpreting tasks, because it may affect coordination and performance without altering the internal acoustic organization of a prerecorded utterance (Moser-Mercer, 2005; Mouzourakis, 2006; Salaets & Brône, 2023).

Playback equipment, presentation level, and listening environment should remain controlled across conditions. Loudness normalization may be used only when it does not remove the property targeted by the manipulation. Any normalization, filtering, or post-processing applied after transmission should be reported.

(3) Acoustic and functional measures

Acoustic modification and functional cue-availability reduction must be evaluated separately.

Acoustic measures should therefore be selected according to a prespecified prosodic function rather than collected as a general inventory of signal differences.

Table 2. Compact operationalization of acoustic measures, independent functional-cue tasks, and downstream outcomes

Prosodic domain	Acoustic metric	Independent functional-cue task	Prespecified downstream outcome
Timing	Local or normalized duration; speech/articulation rate only when relevant	Forced-choice identification of boundary location or a relative-timing contrast in controlled, non-RSI stimuli	Separate message-level listener outcome (intelligibility, comprehensibility, or listening effort); target-text accuracy and adequacy in RSI, including preservation of temporal or grouping relations
Pausing	Number, duration, location, and boundary alignment	Boundary detection or segmentation decision	Comprehensibility or effort; discourse segmentation in RSI
Prominence	Within-item duration, intensity/level, and F_0 contrasts	Prominence identification or contrast-choice task	Comprehensibility or effort; preservation of information structure in RSI
F_0 movement	Range, slope, alignment, and contour measures for the target contrast	Forced-choice identification of a tonal-grouping or pragmatic contrast in controlled, non-RSI stimuli	Separate message-level listener outcome (intelligibility, comprehensibility, or listening effort); target-text accuracy and adequacy in RSI, including preservation of pragmatic relations

Functional-cue tasks must be prevalidated: the target cue and alternative lexical, syntactic, discourse, or visual support should be manipulated independently where feasible, and difficulty, response bias, reliability, and the criterion for meaningful reduction should be prespecified. Matched baseline and transformed signals should be compared within speaker and item. A task is not a valid cue-availability measure if it can be solved through uncontrolled alternative information. Segmental phenomena belong outside the construct unless their interaction with a specified prosodic function is directly operationalized.

(4) Perceptual, effort, and interpreting outcomes

The framework distinguishes four principal outcomes: intelligibility, comprehensibility (Derwing & Munro, 2015; Kang et al., 2025), listening effort (McGarrigle et al., 2014; Pichora-Fuller et al., 2016; Ward et al., 2017), and interpreting performance (Bartłomiejczyk, 2007; Han, 2022). They should not be combined into a single index of speech quality.

Intelligibility should be measured through recognition-based tasks, preferably orthographic transcription or controlled repetition scored according to prespecified units. Word, syllable, or proposition-level scoring may be used depending on the material, but the scoring procedure must be established before analysis.

Comprehensibility concerns perceived ease of understanding and should be measured through a clearly anchored rating scale. It does not indicate how much of the message was correctly recognized and must therefore be reported separately from transcription accuracy.

Listening effort concerns the resources engaged to achieve a given level of performance and should not be inferred from a single index. Subjective ratings, response times, dual-task costs, pupil measures, and other physiological indices may also reflect arousal, motivation, motor demands, or general difficulty; their interpretation and confound controls should therefore be prespecified. Stable intelligibility with increased effort supports a cost–performance dissociation, not compensation unless the independent evidential criteria for compensatory

processing specified in Section 3.4 are also met.

Interpreting performance should be assessed with a prespecified rubric defining the unit of analysis, omissions, additions, distortions, dependent errors, severity weights, and treatment of discourse relations. Delivery and fluency should remain separate from message accuracy. Raters should be blinded and trained using common anchor examples; inter-rater reliability or agreement should be reported with uncertainty intervals. Self-evaluation should remain a separate metacognitive outcome (Bartłomiejczyk, 2007; Han, 2022).

Participants' hearing status, language background, proficiency, interpreting experience, familiarity with the accent, and relevant exposure to the speaker should be documented. Stimulus order and condition order should be randomized or counterbalanced. A listener should not hear the same linguistic item repeatedly in different channel conditions when prior exposure could facilitate recognition; instead, item versions should be distributed across participants through a counterbalanced design.

(5) Staged validation in RSI

Validation should proceed through three linked stages. Stage 1 should test the signal-internal pathway in controlled perception: whether a documented transformation changes the targeted acoustic cue, reduces an independent functional-cue measure, and is associated with prespecified intelligibility, comprehensibility, or effort outcomes. It does not by itself establish an RSI effect.

Stage 2 should examine the RSI pathways separately. Stage 2a tests source-to-interpreter transmission using prespecified target-text outcomes, with lag, effort, and delivery analysed separately; interpreter status, source difficulty, preparation, visual access, topic familiarity, delivery rate, language direction (Chou et al., 2021), and order should be controlled or modelled. Stage 2b treats the interpreter's output as the input signal and tests interpreter-to-audience transmission with independently specified devices, channel processing, listeners, and perceptual outcomes.

Stage 3 should cross the two pathways only after separate validation, estimating whether source-pathway difficulty changes the interpreter's production and whether audience-pathway transformation compounds that change. The design should account for interpreters, speeches/items, audience listeners, transmission runs, order, fatigue, adaptation, and carryover. Transfer from Stage 1 to RSI requires evidence that the cue-function relation remains comparable across stages.

(6) Analytical and reporting requirements

Speakers, items, listeners or interpreters, and transmission runs are distinct sampling units and should be represented in hierarchical models appropriate to the design and outcome distribution (Barr et al., 2013). Channel condition should be controlled and source-signal properties retained as continuous predictors. Critical tests concern transformation-by-source interactions; speakers must not be classified from the same outcomes the model seeks to explain.

Sample-size planning should simulate the intended crossed design and target the smallest theoretically or practically meaningful effect (Brysbaert & Stevens, 2018). The primary manipulation and contrast, prosodic function, acoustic and functional-cue measures, primary and complementary outcomes, equivalence or non-inferiority bounds, exclusions, scoring, and models should be preregistered (Nosek et al., 2018).

Randomizing a channel manipulation supports causal claims only about that manipulation within the tested implementation. Links involving measured source properties, cue availability, or compensation remain associational unless the design identifies mediation; otherwise, use "is associated with" rather than causal language.

Report effect sizes and uncertainty intervals, missingness, transmission failures, exclusions, multiplicity, and deviations. Where consent and licensing permit, share de-identified or controlled-access stimuli, processing scripts,

measurements, scoring materials, and code; because voices may remain identifiable, state residual reidentification risk, access conditions, retention, and reuse restrictions.

(7) Ethical and data-governance requirements

Any future study involving speakers, interpreters, listeners, voice recordings, or physiological measures must obtain ethics approval or a documented exemption before data collection. Protocols should specify consent, withdrawal, compensation, participant burden, secondary use, and whether identifiable voice data may be shared openly, under controlled access, or not at all. Data minimization, encryption, access roles, retention and deletion, licensing, and residual reidentification risk should be prespecified. These are requirements for future implementations; the present conceptual article claims no approval for an unimplemented study.

5. Implications for RSI Research and Professional Practice

Because the framework has not yet been empirically validated, the implications that follow are conditional: they specify how claims about delivery, training, and technology should be evaluated rather than presuming their effectiveness in RSI.

(1) From general audio quality to functionally specified assessment

The model calls for a shift from global assessments of audio quality to functionally specified analysis. In professional discourse, a “connection problem” may refer to several technically and perceptually different events, including delay, signal interruption, bandwidth restriction, dynamic processing, packet loss, or instability in playback level. These events should not be treated as interchangeable, since they may affect different properties of the signal and produce different consequences for interaction and comprehension.

Assessment should therefore link a documented channel transformation to a specific acoustic cue, communicative function, and listener or task outcome; neither a global quality change nor an isolated acoustic difference is sufficient.

Platform comparisons should follow the reporting requirements specified in Section 4.2: general quality ratings or standards compliance alone do not demonstrate preservation of every functionally relevant cue (ISO, 2022; ITU, 1996).

At the operational level, the framework supports practices that improve traceability without presuming a particular prosodic effect. These include pre-event checks using the actual microphones, networks, interfaces, and listening devices involved; documentation of relevant audio settings; verification of the signal received by interpreters rather than only the signal sent; continuous monitoring during the event; and an explicit procedure through which interpreters can report degradation and request technical intervention. Such procedures do not test the model, but they make it possible to identify and document the conditions under which a communication problem occurs.

(2) Delivery adaptation as a testable professional competence

The model also generates hypotheses concerning the role of delivery adaptation in RSI. If specified channel transformations reduce the availability of particular prosodic cues, deliberate control of rate, pausing, prominence, and articulatory definition may help preserve intelligibility. This possibility, however, remains to be empirically evaluated. It should not be converted into a universal recommendation that interpreters speak more slowly, pause more frequently, or produce stronger accentual contrasts in every remote setting. Research on clear speech shows that listener-oriented adaptations are condition-sensitive rather than uniform; speakers alter rate, F_0 , intensity, and articulation differently across adverse conditions (Hazan & Baker, 2011; Smiljanić & Bradlow, 2009). Evidence from

interpreting training similarly suggests that prosody awareness can improve performance, though the existing evidence concerns consecutive interpreting and cannot simply be transferred to RSI (Yenkimaleki & van Heuven, 2018).

Each potential adaptation involves trade-offs. A slower delivery may make local articulation more accessible but increase the interpreter's lag from the source speaker. Longer or more clearly placed pauses may support segmentation but could also disrupt continuity or reduce the time available for subsequent information. Increased prominence may strengthen an intended contrast, yet excessive prominence could alter information structure, pragmatic force, or perceived naturalness. Greater articulatory control may improve segmental accessibility while imposing additional production effort. The effectiveness of an adaptation is therefore likely to depend on the channel transformation, source-speech characteristics, language pair, interpreting strategy, task demands, and individual interpreter.

Training studies should test these adaptations systematically rather than assume their benefits. A suitable design could compare an interpreter's habitual delivery with one or more trained adaptations under documented baseline and transformed conditions. Intelligibility, comprehensibility, listening effort, interpreting accuracy, output fluency, lag, and interpreter fatigue should be evaluated separately. A strategy would count as beneficial only if it improves a relevant outcome without producing a disproportionate cost in another part of the interpreting process.

Under this view, the relevant competence is not a single "digital speaking style" but the ability to make calibrated decisions. Training should combine controlled vocal experimentation with basic technical literacy: recognizing signal instability, distinguishing technical failure from source-speech difficulty, assessing whether a limited delivery adjustment is feasible, and reporting problems when technical intervention is required (Ahrens et al., 2021; Hazan & Baker, 2011; Hlavac, 2013; Smiljanić & Bradlow, 2009; Yenkimaleki & van Heuven, 2018).

Delivery adaptation must not, however, be used to transfer responsibility for inadequate technical conditions to the interpreter. Vocal adjustment cannot compensate reliably for severe packet loss, signal interruption, inappropriate equipment, or failure to meet applicable audio requirements. An important professional competence may sometimes consist precisely in recognizing that adaptation is insufficient and that the technical problem must be corrected before interpreting can continue effectively.

(3) Transmitted vocal delivery as an adaptable and non-normative resource

The broader construct needed here is transmitted vocal delivery, not voice quality in its narrow phonatory or timbral sense. It comprises the production- and channel-dependent properties through which listeners receive linguistic, prosodic, and pragmatic information. Phonatory voice quality, articulation, prosody, intelligibility, and professional delivery should remain analytically distinct (Schiller et al., 2023; Wang, 2022).

Three dimensions may be distinguished for analytical purposes. The first is a production-control dimension, including articulatory coordination, temporal organization, pause placement, prominence, intonational readability, and stability of vocal level. The second is a technical-adaptive dimension, concerning the interpreter's capacity to monitor the transmitted signal and, where appropriate, modify delivery in response to a specified condition. The third is an interactional-pragmatic dimension, concerning the use of voice to convey discourse relations, stance, confidence, turn structure, and communicative continuity when visual or contextual information is limited.

These dimensions do not define a universally optimal professional voice. A vocal feature that supports intelligibility in one language, channel, or task may be unnecessary or counterproductive in another. Nor should the proposed construct be used to classify particular accents, varieties, pitch ranges, or voice types as intrinsically unsuitable for RSI. Prosodic vulnerability belongs to the relation among signal, transformation, listener, and task, not to the personal identity of the speaker. Any professional assessment must consequently distinguish demonstrable functional consequences from preferences or biases concerning accent, gendered voice characteristics, or

conventional expectations of vocal authority.

Voice training for RSI should therefore aim at flexible control rather than normalization, while preserving linguistic identity, pragmatic meaning, and naturalness. Any proposed intervention must be evaluated through the staged agenda in Section 4.

6. Limitations and Future Research

(1) Scope and limitations of the proposed model

The framework has not yet been empirically validated; claims about differential intelligibility, prosodic vulnerability, delivery adaptations, training, or technological benefits therefore remain hypotheses.

The model addresses a specific acoustic-perceptual mechanism rather than the complete ecology of RSI. Remote-interpreting performance also depends on source-text complexity, language pair, topic familiarity, preparation, visual access, interface design, turn management, interpreter experience, fatigue, and organizational conditions (Hale et al., 2022; Moser-Mercer, 2005; Salaets & Brône, 2023).

The model also simplifies the architecture of RSI by representing mediation as a sequence from signal transformation to listener outcome. In actual RSI, at least two transmission pathways must be distinguished. The first carries the source speaker's voice to the interpreter and may affect source-speech comprehension, listening effort, and the formulation of the target text. The second carries the interpreter's output to the audience and may affect the intelligibility and pragmatic accessibility of the interpreted message. The two pathways can involve different devices, processing systems, network conditions, and listeners. Future applications of the model must therefore avoid collapsing source-to-interpreter degradation and interpreter-to-audience degradation into a single channel effect.

A further limitation concerns interactional and multimodal phenomena. Constant latency, for example, may leave the internal acoustic organization of an utterance largely unchanged while affecting turn-taking, coordination, relay interpreting, requests for clarification, and communication with technicians or moderators. Visual access, presentation slides, captions, chat functions, lip movements, and contextual information may compensate for reduced auditory information or introduce additional demands. These effects fall outside a narrowly acoustic account of functional cue availability and require complementary interactional and multimodal analysis. Multimodal effects may also concern the interpreter's own production: restricting gesture has been associated with increased disfluency in trainee consecutive interpreting, although transfer to RSI remains untested (Zhang & Jing, 2025).

Functional cue availability cannot be inferred directly from acoustic change; it must be estimated through targeted listener tasks. The resulting estimate will depend partly on how the relevant linguistic function and behavioural measure are operationalized.

Compensatory processing presents a related difficulty. Greater reliance on contextual, lexical, syntactic, or visual information may preserve recognition despite reduced acoustic evidence, but compensation is not directly observable through intelligibility scores alone. Subjective effort ratings provide useful information but may be affected by individual response strategies. Behavioural and physiological measures introduce their own interpretive limitations. The model consequently requires converging evidence rather than treating any single measure as a transparent representation of listener effort (McGarrigle et al., 2014; Pichora-Fuller et al., 2016).

The generalizability of the model also remains undetermined. Cue functions and cue weighting vary across languages, varieties, accents, speakers, and communicative tasks. Listener familiarity, proficiency, hearing ability, professional experience, and contextual knowledge may further influence whether a transformation becomes consequential. Evidence obtained from one language pair, voice type, or controlled listening task should therefore

not be generalized automatically to other RSI settings.

Adaptive or proprietary processing limits reproducibility because codecs, noise suppression, automatic gain control, and packet-loss concealment may vary across software versions, devices, and network states (Raake, 2006). Claims must therefore be restricted to the documented implementation and testing conditions.

(2) Priorities for empirical validation and generalization

The first research priority is to test the complete evidential pathway through independent measures of each link, as specified in Sections 3.4 and 4.

Studies should independently measure recognition, comprehensibility, and effort to determine when compensation preserves accuracy and when it becomes insufficient as task demands increase (Peelle, 2018; Pichora-Fuller et al., 2016; Ward et al., 2017; Wild et al., 2012).

A second priority concerns the boundary conditions of differential vulnerability. Research should examine whether salience and cue redundancy predict resilience across different prosodic functions (Aylett & Turk, 2004; Wagner & Watson, 2010). A boundary marked simultaneously by pausing, final lengthening, and tonal movement may remain accessible when one dimension is weakened, whereas a contrast depending principally on a single subtle cue may be more easily disrupted. Such predictions should be evaluated across languages and varieties whose prosodic systems assign different weights to duration, pitch, intensity, and rhythmic organization.

Speaker and listener diversity must form part of this validation rather than being treated as uncontrolled variation (Kang et al., 2025; Schiller et al., 2023). Studies should include multiple speakers with continuously measured signal properties and listeners differing in language proficiency, interpreting experience, accent familiarity, and relevant hearing characteristics. This would make it possible to determine whether an observed effect reflects a general channel–cue interaction, a listener-specific difficulty, or familiarity with a particular speaker or variety.

Because existing studies on interpreted prosody, prosodic processing, and prosody-awareness training do not validate the complete channel–cue–outcome sequence in RSI (Du & Wang, 2025; Lenglet & Michaux, 2020; Yenkimaleki & van Heuven, 2018), controlled findings should subsequently be tested in longer discourse and simultaneous-interpreting tasks, including their effects on retention, reformulation, target-text accuracy, delivery, and effort.

The two RSI transmission pathways should first be studied separately and then factorially, to determine whether difficulty in source-to-interpreter transmission changes the interpreter's production and compounds degradation in interpreter-to-audience transmission.

Interactional delay and multimodal compensation constitute an additional research strand (Moser-Mercer, 2005; Salaets & Brône, 2023). Latency should be investigated through tasks involving live coordination rather than treated as an acoustic modification of prerecorded material. Similarly, studies should determine when video, captions, presentation materials, or contextual information compensate for reduced auditory cues and when managing these additional resources increases rather than reduces cognitive demand.

Longitudinal studies should determine whether interpreters can learn to recognize relevant channel conditions and apply beneficial adaptations selectively without increasing lag, fatigue, information loss, or pragmatic distortion.

(3) Towards prosody-aware tools

Prosody-aware technology should be treated as a research objective rather than as an already defined product category. The term should refer only to systems whose design and evaluation explicitly consider the preservation or accessibility of functionally relevant prosodic cues. It should not imply that an algorithm can reconstruct a universally correct prosodic form or determine communicative meaning from acoustic measures alone. Self-supervised speech models can learn rich representations directly from raw audio (Baevski et al., 2020), but

representation learning alone does not validate a system as prosody-aware in the functional sense proposed here.

Four possible technological functions can be distinguished.

Table 3. Possible functions and evidential requirements of prosody-aware tools

Function	Intended purpose	Minimum evidence required	Principal risk
Diagnostic	Detect transformations that may reduce relevant cue availability	Validated relation among signal measure, functional cue, and listener outcome	Treating an acoustic proxy as certain evidence of communication failure
Preservation	Limit avoidable modification of relevant contrasts during transmission	Matched input–output analysis and demonstrated perceptual benefit	Introducing artifacts or protecting one cue at the expense of another
Support	Provide interpreters or technicians with actionable information about channel conditions	Evidence that feedback improves intervention without increasing workload excessively	Adding distraction, false alarms, or misplaced responsibility
Enhancement	Modify a transmitted signal to increase the accessibility of selected cues	Demonstrated benefit for intelligibility or effort together with preservation of pragmatic meaning	Altering information structure, vocal identity, naturalness, or speaker intention

Diagnostic systems represent the most cautious initial direction. Where matched input and output signals are available, such systems could identify changes in temporal, tonal, or dynamic properties associated with a specified function. In live conditions without a clean reference signal, they could monitor technically identifiable events or unusual signal behaviour. Their output should nevertheless be expressed probabilistically: detecting a change in an acoustic parameter does not demonstrate that listeners have lost access to the corresponding function.

Preservation-oriented tools could seek to prevent unnecessary reduction of contrasts during encoding, dynamic processing, or transmission. Their objective would not be to amplify prosody indiscriminately, but to retain distinctions whose perceptual value has been established. Any benefit must be demonstrated through listener outcomes rather than inferred from acoustic similarity to the input signal. A technically faithful signal is not necessarily functionally sufficient, and a measurable acoustic difference is not necessarily harmful.

Support tools could provide interpreters or technicians with warnings, visualizations, or comparative indicators when channel conditions change. Their practical value would depend on whether the information is understandable, timely, and connected to an available corrective action. Feedback that cannot guide intervention, generates frequent false alarms, or competes with the interpreter's attention could increase rather than reduce workload.

Enhancement systems involve the greatest methodological and ethical risk because they actively modify transmitted speech. Reinforcing prominence, altering temporal relations, reconstructing contours, or changing dynamic contrasts may affect information structure, stance, naturalness, and perceived identity. Such processing should therefore remain transparent, reversible where possible, and subject to human control. It should not normalize voices according to a single accentual, gendered, or culturally dominant model.

Evaluation of these tools must extend beyond global audio quality and user preference. Relevant criteria include intelligibility, comprehensibility, listening effort, interpreting accuracy, artifact introduction, pragmatic preservation, vocal naturalness, system reliability, and the workload created by any feedback interface. Testing should also examine whether performance differs across languages, accents, voice characteristics, devices, and network conditions. A tool that benefits one narrowly defined speech profile while degrading others would not constitute a general solution.

The most realistic near-term objective is therefore not automatic reconstruction of prosody, but improved measurement, preservation, and transparent reporting of conditions that may reduce the accessibility of relevant cues. Technology should support professional judgment and technical intervention without substituting for the interpreter's competence or obscuring responsibility for inadequate transmission conditions.

7. Conclusion

This article has argued that digital mediation in RSI cannot be understood through general references to poor connection or audio degradation alone. The framework separates source-signal organization, channel transformation, acoustic cue modification, functional cue availability, compensatory processing, and observable outcomes.

Prosodic vulnerability is the estimated susceptibility of a functionally relevant prosodic contrast, for a specified source signal, listener population, task, channel transformation, and baseline, to reduced functional cue availability. It is not an intrinsic property of a speaker, accent, variety, pitch range, or voice type. Acoustic change, cue-availability reduction, effort, and task performance are distinct variables; no one alone establishes the proposed mechanism.

Source-to-interpreter and interpreter-to-audience pathways require separate validation before their interaction is tested. If supported, specific channel-cue relations may inform functionally specified assessment, calibrated delivery adaptation, and prosody-aware tools. This article advances the normative position that responsibility for intelligibility in RSI should be treated as distributed across speakers, interpreters, equipment, processing systems, networks, organizers, and listening environments. This position is conceptually consistent with the framework but is not presented as an empirical finding or as a requirement derived from professional standards.

The contribution is a set of distinctions, falsification criteria, and reporting requirements whose value depends on whether future evidence confirms, narrows, or rejects the proposed relations.

References

- [1] Ahrens, B. (2005). Prosodic phenomena in simultaneous interpreting: A conceptual approach and its practical application. *Interpreting*, 7(1), 51–76. <https://doi.org/10.1075/intp.7.1.04ahr>.
- [2] Ahrens, B., Beaton-Thome, M., & Rütten, A. (2021). The pivot to remote online teaching on the MA in Conference Interpreting in Cologne: Lessons learned from an unexpected experience. *The Journal of Specialised Translation*, 36, 251–284. <https://doi.org/10.26034/cm.jostrans.2021.043>.
- [3] Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1), 31–56. <https://doi.org/10.1177/00238309040470010201>.
- [4] Baeovski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 12449–12460). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html>.
- [5] Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>.
- [6] Bartłomiejczyk, M. (2007). Interpreting quality as perceived by trainee interpreters: Self-evaluation. *The Interpreter and Translator Trainer*, 1(2), 247–267. <https://doi.org/10.1080/1750399X.2007.10798760>.
- [7] Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1(1), Article 9. <https://doi.org/10.5334/joc.10>.
- [8] Chakhachiro, R. (2016). Contribution of prosodic and paralinguistic cues to the translation of evidentiary audio recordings. *Translation & Interpreting*, 8(2), 46–63. <https://doi.org/10.12807/ti.108202.2016.a04>.
- [9] Cheung, A. K. F. (2024). Cognitive load in remote simultaneous interpreting: Place name translation in two Mandarin variants. *Humanities and Social Sciences Communications*, 11, Article 1238. <https://doi.org/10.1057/s41599-024-03767-y>.
- [10] Chmiel, A., Spinolo, N., Korpala, P., Olalla-Soler, C., Rozkrut, P., Kajzer-Wietrzny, M., & Ghiselli, S. (2025). The impact of remote interpreting settings on interpreter experience and performance. *Translation and Interpreting Studies*, 20(2), 212–243. <https://doi.org/10.1075/tis.24125.chm>.
- [11] Chou, I., Liu, K., & Zhao, N. (2021). Effects of directionality on interpreting performance: Evidence from interpreting between Chinese and English by trainee interpreters. *Frontiers in Psychology*, 12, Article 781610. <https://doi.org/10.3389/fpsyg.2021.781610>.
- [12] Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40(2), 141–201. <https://doi.org/10.1177/002383099704000203>.

- [13]Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins. <https://doi.org/10.1075/llt.42>.
- [14]Du, M., & Wang, B. (2025). How are prosodic features processed cognitively? A case study of the multimodal cognitive dynamics in simultaneous interpreting with text. *Meta*, 70(1), 194–220. <https://doi.org/10.7202/1125408ar>.
- [15]Ershaid, H., Lizarazu, M., McLaughlin, D., Cooke, M., Simantiraki, O., Koutsogiannaki, M., & Lallier, M. (2024). Contributions of listening effort and intelligibility to cortical tracking of speech in adverse listening conditions. *Cortex*, 172, 54–71. <https://doi.org/10.1016/j.cortex.2023.11.018>.
- [16]Gile, D. (2009). *Basic concepts and models for interpreter and translator training* (Rev. ed.). John Benjamins. <https://doi.org/10.1075/btl.8>.
- [17]Hale, S. B., Goodman-Delahunty, J., Martschuk, N., & Lim, J. (2022). Does interpreter location make a difference? A study of remote vs face-to-face *interpreting* in simulated police interviews. *Interpreting*, 24(2), 221–253. <https://doi.org/10.1075/intp.00077.hal>.
- [18]Han, C. (2022). Interpreting testing and assessment: A state-of-the-art review. *Language Testing*, 39(1), 30–55. <https://doi.org/10.1177/02655322211036100>.
- [19]Hazan, V., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, 130(4), 2139–2152. <https://doi.org/10.1121/1.3623753>.
- [20]Hlavac, J. (2013). Should interpreters be trained and tested in telephone and video-link interpreting? Responses from practitioners and examiners. *International Journal of Interpreter Education*, 5(1), 34–50. <https://open.clemson.edu/ijie/vol5/iss1/4/>.
- [21]Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, 119(5), 3059–3071. <https://doi.org/10.1121/1.2188377>.
- [22]International Organization for Standardization. (2022). *Simultaneous interpreting delivery platforms—Requirements and recommendations* (ISO Standard No. 24019:2022). <https://www.iso.org/standard/80761.html>.
- [23]International Organization for Standardization. (2025). *Simultaneous interpreting—Equipment—Requirements* (ISO Standard No. 20109:2025). <https://www.iso.org/standard/84056.html>.
- [24]International Telecommunication Union. (1996). *Methods for subjective determination of transmission quality* (ITU-T Recommendation P.800). <https://www.itu.int/rec/t-rec-p.800-199608-i>.
- [25]Kang, O., Bae, Y., & Miao, Y. (2025). L2 speech intelligibility. *Language Teaching*. Advance online publication. <https://doi.org/10.1017/S0261444825101006>.
- [26]Lenglet, C., & Michaux, C. (2020). The impact of simultaneous-interpreting prosody on comprehension: An experiment. *Interpreting*, 22(1), 1–34. <https://doi.org/10.1075/intp.00032.len>.
- [27]Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7–8), 953–978. <https://doi.org/10.1080/01690965.2012.705006>.
- [28]McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group “white paper.” *International Journal of Audiology*, 53(7), 433–440. <https://doi.org/10.3109/14992027.2014.890296>.
- [29]Moser-Mercer, B. (2005). Remote interpreting: Issues of multi-sensory integration in a multilingual task. *Meta*, 50(2), 727–738. <https://doi.org/10.7202/011014ar>.
- [30]Mouzourakis, P. (2006). Remote interpreting: A technical perspective on recent experiments. *Interpreting*, 8(1), 45–66. <https://doi.org/10.1075/intp.8.1.04mou>.
- [31]Niebuhr, O., & Siegert, I. (2023). A digital “flat affect”? Popular speech compression codecs and their effects on emotional prosody. *Frontiers in Communication*, 8, Article 972182. <https://doi.org/10.3389/fcomm.2023.972182>.
- [32]Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>.
- [33]Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, 39(2), 204–214. <https://doi.org/10.1097/AUD.0000000000000494>.
- [34]Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing*, 37(Suppl. 1), 5S–27S. <https://doi.org/10.1097/AUD.0000000000000312>.
- [35]Raake, A. (2006). *Speech quality of VoIP: Assessment and prediction*. John Wiley & Sons. <https://doi.org/10.1002/9780470033005>.
- [36]Rönnerberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, 7, Article 31. <https://doi.org/10.3389/fnsys.2013.00031>.
- [37]Roziner, I., & Shlesinger, M. (2010). Much ado about something remote: Stress and performance in remote interpreting.

- Interpreting*, 12(2), 214–247. <https://doi.org/10.1075/intp.12.2.05roz>.
- [38]Salaets, H., & Brône, G. (2023). “Working at a distance from everybody”: Challenges (and some advantages) in working with video-based interpreting platforms. *The Interpreters’ Newsletter*, 28, 189–209. <https://doi.org/10.13137/2421-714X/35556>.
- [39]Schiller, I. S., Aspöck, L., & Schlittmeier, S. J. (2023). The impact of a speaker’s voice quality on auditory perception and cognition: A behavioral and subjective approach. *Frontiers in Psychology*, 14, Article 1243249. <https://doi.org/10.3389/fpsyg.2023.1243249>.
- [40]Seeber, K. G. (2011). Cognitive load in simultaneous *interpreting*: Existing theories—New models. *Interpreting*, 13(2), 176–204. <https://doi.org/10.1075/intp.13.2.02see>.
- [41]Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass*, 3(1), 236–264. <https://doi.org/10.1111/j.1749-818X.2008.00112.x>.
- [42]Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7–9), 905–945. <https://doi.org/10.1080/01690961003589492>.
- [43]Wang, C. (2022). A theoretical model to elucidate the elusive concept ‘voice’ for interpreters. *Perspectives*, 30(4), 569–584. <https://doi.org/10.1080/0907676X.2021.1922472>.
- [44]Ward, K. M., Shen, J., Souza, P. E., & Grieco-Calub, T. M. (2017). Age-related differences in listening effort during degraded speech recognition. *Ear and Hearing*, 38(1), 74–84. <https://doi.org/10.1097/AUD.0000000000000355>.
- [45]Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful listening: The processing of degraded speech depends critically on attention. *The Journal of Neuroscience*, 32(40), 14010–14021. <https://doi.org/10.1523/JNEUROSCI.1528-12.2012>.
- [46]Yenkimaleki, M., & van Heuven, V. J. (2018). The effect of teaching prosody awareness on interpreting performance: An experimental study of consecutive interpreting from English into Farsi. *Perspectives*, 26(1), 84–99. <https://doi.org/10.1080/0907676X.2017.1315824>.
- [47]Zhang, Q., & Jing, Y. (2025). The impact of interpreting students’ gestures and speech content on speech fluency of consecutive interpreting. *Frontiers in Psychology*, 16, Article 1568341. <https://doi.org/10.3389/fpsyg.2025.1568341>.
- [48]Zhu, X., & Aryadoust, V. (2022). A synthetic review of cognitive load in distance interpreting: Toward an explanatory model. *Frontiers in Psychology*, 13, Article 899718. <https://doi.org/10.3389/fpsyg.2022.899718>.