Time Blindness in STI Intelligibility Testing

Why STIPA misses the transient half of speech, and the early reflections that destroy it

An SDA Consulting White Paper

Author: Robert Oswood, SDA Consulting · roswood@sda-consulting.com

Executive Summary

STIPA, the field method for the Speech Transmission Index in IEC 60268-16, is blind to speech transients. Research by Yoo and colleagues found that transient components carry about 2% of speech energy yet are nearly as intelligible as the full signal. STIPA is built on slow energy envelopes, so its result is dominated by the quasi-steady-state energy and largely ignores the transient information listeners depend on.

The consequence matters most when listeners cannot lean on context: unfamiliar words, names, place names, codes and emergency instructions. A system can pass STI ≥ 0.50 while the transient cues those messages depend on are being smeared. Two secondary mechanisms make this worse and also escape STIPA: strong early reflections at 7 to 10 ms, and discrete late echoes.

1. What STIPA Actually Measures

STIPA measures how well slow fluctuations in speech energy survive the transmission path. Its test signal is stationary noise, shaped to a speech spectrum and amplitude-modulated at two frequencies per octave band. Across all seven bands (125 Hz to 8 kHz), those modulation frequencies run from 0.63 Hz to 12.5 Hz.

Three features of this design matter here:

- **Envelope only.** The metric reflects how much modulation depth is lost, averaged over time. It does not see the fine timing of events within a syllable.
- **Slow modulations only.** The fastest modulation tested, 12.5 Hz, has an 80 ms period. Many speech transients, such as stop bursts and rapid formant transitions, unfold over a few tens of milliseconds.
- **No transients in the stimulus.** The STIPA signal has no onsets, bursts or glides. The test never sends the kind of event whose degradation it would need to detect.

2. The Primary Problem: Speech Transients

2.1 The evidence

Yoo, Boston and colleagues at the University of Pittsburgh split high-pass-filtered speech into quasi-steady-state (QSS) and transient components using time-varying filters. The (QSS) component includes both vowels and consonants; transients include onsets, offsets and transitions such as vowel-to-consonant movements.

Their findings:

- The transient component held only about **2% of the speech energy**.
- Yet it contributed **almost as much to the intelligibility of speech as the** (QSS) **component with 98% of speech energy.**
- Amplifying transients relative to (QSS) improved intelligibility in noise.

2.2 Why STIPA cannot see it

STI weights intelligibility by energy-based modulation transfer. When 98% of the energy sits in the QSS component, the index effectively measures how that component is transmitted. Damage confined to the low-energy transient component barely moves the result, even though, on Yoo's evidence, that component carries much of what makes speech intelligible.

Put simply, STIPA assesses the part of speech that is loud, but listeners also rely heavily on the part that is quiet and brief.

2.3 Familiar versus unfamiliar material

The author's position is that this gap is masked in everyday listening. Common messages draw on a vocabulary of a few hundred frequent words, and listeners fill gaps from context and expectation. Steady-state cues plus redundancy are often enough, so STI correlates well with intelligibility for this material.

Unfamiliar material removes that support. Proper names, place names, platform numbers, flight codes and non-routine emergency instructions must be decoded from the acoustic signal itself. Distinguishing "gate B13" from "gate D30" depends on consonant bursts and transitions, precisely the transient cues STIPA does not test. This is where a passing STI result is least trustworthy and where the consequences are most serious.

3. Secondary Problem: Early Reflections at 7 to 10 ms

A strong reflection arriving 7 to 10 ms after the direct sound causes comb filtering, with notches spaced about 100 to 140 Hz apart across the speech band. These notches strip spectral detail from formants and consonant cues, and the overlapping copy smears the onsets of transients.

STIPA misses this for two reasons:

- **Octave-band averaging.** Narrow notches at 100 to 140 Hz spacing average out within each octave band.
- **Short delay versus slow modulation.** A 10 ms delay barely affects modulation up to 12.5 Hz. In a simple model with a reflection 3 dB below the direct sound, the modulation transfer at 12.5 Hz stays above 0.9.

The result is a high STI value for a path that audibly damages the transient detail described in Section 2.

3.1 Phase coherence and proximity: Griesinger's work

David Griesinger's research gives a perceptual explanation for why early reflections harm intelligibility more than STI suggests. He argues that the ear extracts much of its information from the phase alignment of upper harmonics, chiefly in the 1 to 4 kHz range. When the direct sound arrives with those

harmonics coherent, the talker sounds close and distinct, a quality he calls proximity. It makes speech easier to follow, to localize and to separate from other sources.

Reflections arriving within roughly the first 100 ms scramble that phase relationship before the ear can decode it. The talker then sounds distant and muddy, and listeners need more effort to understand, even when the words are technically audible. Griesinger proposed an impulse response measure, LOC (localizability), to predict whether the direct sound will be perceived as distinct against the early reflected energy. Working in the 1 to 4 kHz range, LOC compares the level of the direct sound (the first 5 ms) with the average logarithmic build-up of reflected energy over a window of about 100 ms, ignoring reflections more than about 20 dB down. A result of about +3 dB or more indicates the talker will be localized and perceived as close; MATLAB code for the calculation is published on his website.

Griesinger's work shows this mechanism directly. Vocal harmonics start in phase at each glottal pulse, drift apart, and realign once every fundamental period. Above about 1 kHz this produces a periodic peak in the amplitude envelope at the fundamental frequency, roughly 100 to 350 Hz for speech. Griesinger describes an 80 ms autocorrelator in the auditory system that detects this periodicity: proximate sounds show strong, regular autocorrelation peaks at the fundamental, while sounds whose harmonic phases have been randomized by early reflections show weak or absent peaks and are heard as distant.

The contrast with STI is stark. STI examines envelope modulations no faster than 12.5 Hz. The envelope periodicity Griesinger's autocorrelator detects runs at the fundamental, at least ten times faster, and lies entirely outside what STI measures.

This connects directly to the argument of this paper. Phase coherence is carried by fine temporal structure, the same domain as speech transients, and STI cannot see either. A space can score well on STI while destroying the coherence that gives speech proximity, and with it, much of the listener's ability to decode unfamiliar material.

4. Secondary Problem: Late Discrete Echoes

Discrete echoes and late specular reflections from glazing, concrete, stadium roofs and tunnel walls also escape proper penalty. For a single echo, modulation transfer behaves like a comb filter in the modulation domain. Where the echo delay is a whole number of modulation periods, it reinforces the envelope instead of degrading it.

A simple model (single echo 3 dB below the direct sound, no noise, equal band weighting) gives:

Echo delay (ms)	Modelled STI
50	0.78
100	0.74
200	0.66
300	0.66
500	0.69

A strong, clearly audible echo at 500 ms still yields a "good" STI, and the index rises with delay past 300 ms. In this model, full STI and STIPA gave nearly identical values, so the weakness lies in the envelope approach itself, not in STIPA's reduced sampling.

5. Consequences for Consultants

- **False assurance.** A compliant STIPA result may be accepted for a voice alarm system whose unfamiliar or critical messages are hard to decode.
- **Wrong remedies.** The index does not identify transient smearing, early comb filtering or echoes as causes. Absorption or level changes may be specified where re-aiming, delay alignment or treating one surface is the real fix.
- **Unexplained complaints.** Users can report poor intelligibility in spaces that measured well.

The error runs in both directions when STI is the certification requirement for a space.

- **False rejection.** A system that delivers strong, coherent direct sound can preserve enough transient and phase information for listeners to understand speech well, including unfamiliar material. If reverberation or noise pulls the STI below the target, the space still fails. STI cannot recognize that the direct sound carries what listeners need, so a system that works may be rejected, redesigned or expensively treated without need. Griesinger's work supports this. In his paper on the audibility of direct sound as a key to clarity, he argues that clarity depends on whether the direct sound can be detected as separate from the reflections that follow, not on the total ratio of early to late energy. Once the phase-coherent onset of the direct sound is audible, listeners can separate and decode speech in rooms that energy-based measures rate poorly. His LOC measure is designed to predict exactly this threshold from a binaural impulse response, and could serve as a supporting criterion when a well-designed system narrowly fails on STI.

6. Recommended Practice

STIPA remains a useful baseline. These checks fill its blind spots.

1. **Test with unfamiliar material.** Supplement STIPA with listener tests using phonetically balanced or nonsense-syllable material, such as Modified Rhyme Test or CVC word lists, which depend on transient cues. Add jury-based tests with running speech, such as the Connected Speech Test (CST) developed by Cox and colleagues at the University of Memphis Hearing Aid Research Laboratory. The two test types are complementary. Rhyme and nonsense-syllable tests isolate transient-dependent word recognition but strip out prosody. The CST, played to a listener panel in the actual space, preserves the natural rhythm, stress and intonation of connected speech. STIPA assesses neither transients nor prosody.
2. **Inspect the impulse response.** Review the energy-time curve at each position for strong arrivals in the 5 to 15 ms range and isolated peaks beyond 50 ms.
3. **Look at the fine spectrum.** Use narrowband or 1/12-octave transfer function measurements to reveal comb filtering hidden by octave averaging.
4. **Apply an explicit echo criterion**, such as the Dietsch and Kraak criterion, alongside STI.
5. **Design for transients.** Control early reflections near loudspeakers and listeners, set delays between zones, and use directional devices to limit hard-surface returns.

6. **Employ loudspeakers which have superior transient response**. For example (ribbons and AMTs).
7. **Report the limitation.** State in commissioning reports that STIPA does not assess transient transmission and record any supplementary tests.

7. Directions for Standards and Research

- A transient-sensitive test signal or companion metric to sit alongside STIPA.
- Listener studies comparing STI with intelligibility for familiar versus unfamiliar material in the same spaces.
- Quantifying how early reflections and echoes degrade the transient component specifically.
- Normative guidance on supplementary testing for voice alarm systems carrying non-routine messages.

8. Conclusion

STIPA measures the transmission of the energetic, steady part of speech. Yoo's work shows that the brief, low-energy transient part carries intelligibility comparable to 50% of the whole signal. STIPA does not test that part, and it also overlooks early reflections and echoes that damage it. For familiar messages the gap is often hidden by context. For unfamiliar and safety-critical messages it can be decisive. Consultants should treat a passing STIPA result as necessary but not always sufficient.

References

- IEC 60268-16, *Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index*.
- Yoo, S., Boston, J. R., Durrant, J. D., et al., Relative energy and intelligibility of transient speech components, EUSIPCO 2004.
- Yoo, S. et al., Speech signal modification to increase intelligibility in noisy environments, J. Acoust. Soc. Am. 122(2), 2007.
- Houtgast, T. and Steeneken, H. J. M., foundational work on the STI and the modulation transfer function.
- Dietsch, L. and Kraak, W., objective criterion for echo disturbance.
- National voice alarm standards, for example BS 5839-8 and NFPA 72.
- Griesinger, D., The physics of auditory proximity: Auditory mechanisms that extract pitched signals from noise and other signals, J. Acoust. Soc. Am.
- Griesinger, D., Pitch Coherence as a Measure of Apparent Distance.
- Griesinger, D., The audibility of direct sound as a key to measuring the clarity of speech and music.
- Griesinger, D., Acoustic Quality, Proximity, and Localization in Concert Halls, Psychomusicology: Music, Mind & Brain.
- Griesinger, D., MATLAB code for calculating LOC.

© 2026 SDA Consulting. All rights reserved.