



Technical Document

AES TD1009.0.0.1

Improving Dialogue Intelligibility in Media

25 November 2025



Audio Engineering Society, Inc.
697 3rd Avenue, Suite 405
New York
NY 10017
USA

Tel: +1 212 661 8528. Email: hq@aes.org. Internet: <http://www.aes.org>



The Audio Engineering Society's Technical Council and its Technical Committees respond to the interests of the membership by providing technical information at an appropriate level via conferences, conventions, workshops, and publications. They work on developing tutorial information of practical use to the members and concentrate on tracking and reporting the very latest advances in technology and applications. This activity is under the direction of the AES Technical Council and its Committees.

The Technical Council and its first Technical Committees were founded by the Audio Engineering Society in 1979, and standing rules covering their activities were established in 1986, with the intention of defining and consolidating the technical leadership of the Society for the benefit of the membership. The Technical Council consists of the officers of the Technical Council, the chairs of the Technical Committees, the editor of the Journal, and as ex-officio members without vote, the other officers of the Society.

TECHNICAL COUNCIL

Rafael Kassier—Chair

Lesley Fogle—Co-Chair

Jayant Datta—Vice Chair

TECHNICAL COMMITTEES

www.aes.org/technical

The goal of AES Technical Documents is not to state a policy or support for a technical method, device or principle, but to provide information that represents the collective knowledge of a group of experts. Neither the AES nor any of its Committees or members shall be responsible for any consequences resulting from the use of information contained in this publication.

AES Technical Committee on Broadcast and Online Delivery

Co-Chairs: David Bialik, Kimio Hamasaki, Matthieu Parmentier, Jim Starzynski
Study Group on Dialogue Intelligibility

Improving Dialogue Intelligibility in Media

Version 1.0: 25 November 2025

Chair: Jim Starzynski

Editors: Jim Starzynski, Robert Orban, Robert Bleidt

Contributing Authors: Robert Bleidt, Jon Greasley, Hannes Muesch
with: Scott Norcross, Robert Orban, Jim Starzynski

Working Group:

Derek Barrentine
Hannah Baumgartner
David Bialik
Robert Bleidt
Manuel Briand
Richard Friedel
Jon Greasley
Rainier Huber
Scott Isabelle
Mark Johnson
Bob Katz
Scott Kramer
John Kean
Scott Levine
Steve Morris
Hannes Muesch
Scott Norcross
Robert Orban
Matthieu Parmentier
Sean Richardson
Jim Starzynski
Martin Walsh

Table of Contents

1	Document Conventions	1
2	Introduction	1
	Figure 1. The Dialogue Intelligibility Ecosystem	2
3	Human Hearing and Factors Affecting Speech Intelligibility	3
3.1	Bottom-Up and Top-Down Processing.....	4
3.2	Intelligibility versus Listening Effort.....	4
3.3	Factors Impacting Intelligibility.....	5
3.3.1	Speaker and Message.....	5
3.3.2	Degradation of the Speech Signal Through Acoustic and Electronic Transmission	5
3.3.3	Ratio Between Speech Level and Level of Interfering Sounds.....	6
	Figure 2. The proportion of speech cues available to the listener (Level Weight) as a function of speech level. This relationship is described by the Speech Intelligibility Index.	6
3.3.4	The Listener	7
4	Production and Post-Production	8
4.1	The Current State of Production Mixing, Monitoring, and Post-Production.....	8
4.2	Practices for Improvement: Methods for Ensuring Intelligibility	9
4.2.1	Intelligibility in Stereo Listening.....	9
4.2.2	Playback Levels.....	10
4.2.3	Loudness Measurement and Loudness Ranges	10
	Figure 3. Equal-loudness contours (red) (from ISO 226:2003 revision). Original ISO standard shown (blue) for 40 phons.	11
4.2.4	Dialogue Processing in Production and Mixing.....	12
4.2.5	Distribution Tools Relevant to Content Creators.....	13
4.2.6	A Culture of Intelligibility.....	13
	Figure 4. Signal Flow from Mixing through Distribution/Transmission. Suggested practices in yellow.	13
	Figure 5. Signal Flow at the Consumer Destination. Suggested practices in yellow.	14
5	Distribution	15
5.1	Distribution Factors.....	15
5.2	The need for Loudness Normalization and DRC	15
	Figure 6. Loudness Normalization Process.....	16
5.3	Loudness Normalization	16
5.4	Dynamic Range Control.....	17
	Figure 7. Dynamic Range Control Process	18
5.5	Real-time Audio Processing	18

5.6	Downmix in Distribution	19
5.6.1	The need for downmixing	19
	Figure 8. AC-3 Downmixing for Stereo Playback	20
5.6.2	Over-the-Top Television Soundtrack to device matching	21
5.6.3	Metadata Control of the Stereo Downmix Method Used by Decoders	21
5.7	Emerging Options	21
5.7.1	ANSI/CTA-2075-Compliant DRC and Output Levels	21
5.7.2	Distribution technologies to improve intelligibility	22
5.7.3	Algorithms to Estimate Intelligibility	23
6	Consumer Playback Devices and Listening Environments	23
6.1	Televisions	23
	Figure 9. Speaker system of typical TV	24
6.2	Upgraded Audio Systems	25
	Figure 10. Acoustic frequency response of the best-sounding 2023 TV (green) and representative poor-sounding TV (blue) from consumer test website	25
6.3	Device Options for Improving Intelligibility	26
6.3.1	DRC	26
6.3.2	A Distributor’s Alternate Stream or NGA CODEC	26
6.3.3	Post Processing	26
	Figure 11. Increasing Dialogue Intelligibility with DRC and Dialogue Boosting	26
6.3.4	Device Configuration Options	27
6.3.5	Bitstream Passthrough	27
	Figure 12. Audio signal paths from a typical TV’s audio decoders to its speakers and a connected soundbar or AVR. Note that signal (a), typically carried over HDMI, may be stereo PCM, an encoded bitstream, or an uncompressed bitstream	28
6.4	Room Acoustics	28
6.5	Mobile Devices	29
6.6	Loudness Management in Playback	29
7	Annex A: Terms, Definitions, and Abbreviations	31
8	Annex B: Related Loudness Standards and Recommendations	34
9	Annex C: Press Accounts	35

Improving Dialogue Intelligibility in Media

1 Document Conventions

- Underlined text denotes an active hyperlink to another section of this document, often to a definition in Annex A: Terms, Definitions, and Abbreviations or a reference in the following annexes.
- LKFS and LUFS are identical units of measurement specified in ITU-R BS.1770. This document uses LKFS (Loudness K-weighted, relative to digital Full Scale).
- This document uses the term Distributor to refer to the entity in the delivery chain tasked with normalization of content loudness before transmission, streaming or download. This entity is typically a broadcaster, branded streaming service, or on-demand content provider.

For casual readers or professionals who want a quick reference, this document identifies **KEY POINTS**. These are the results of thousands of hours of real world/practical experience.

2 Introduction

This report addresses dialogue intelligibility and the factors that impact the audience's experience. It is intended to guide removal of the barriers impeding fans from fully immersing themselves in the programs they enjoy. Key points throughout the document summarize this guidance.

Key contributors to dialogue intelligibility are addressed:

- Human Hearing
- Content Mixing, Distribution, and Delivery
- Consumer Sound System Capability

This paper is for:

- Content Production Professionals
- Content Distributors
- Device Manufacturers
- Enthusiasts

The Dialogue Intelligibility Ecosystem shown in Figure 1 depicts issues that can arise as content moves from production to distribution to the consumer.

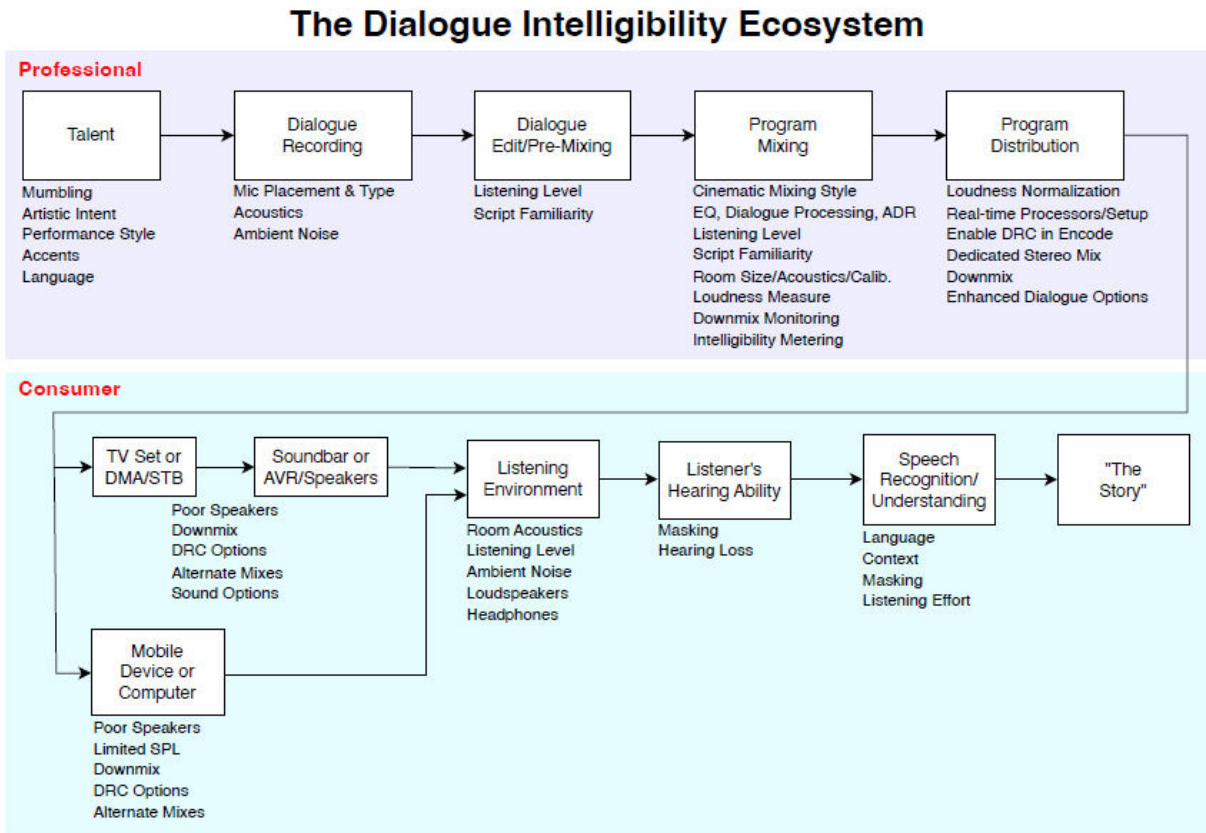


Figure 1. The Dialogue Intelligibility Ecosystem

The best viewing experience is enjoying a great story that does not require significant listening effort. Yet the publications in [Annex C](#) indicate that viewers at home strain to understand dialogue for many reasons including:

- Speech may not be articulated clearly
- Accents can be tough to understand
- Dense soundtracks mask dialogue
- Content dynamic range is excessive
- Noise in the listening environment masks dialogue
- Non-optimal reproduction
- Hearing impairment

These issues, among others, make following the storyline difficult. Dialogue intelligibility is a complex and nuanced issue. Dialogue intelligibility is defined as the ease with which spoken communication can be understood and plays a fundamental role in shaping the audience's experience. Understanding these issues is important to making the best decisions when preparing, distributing, or playing back content.

Stories in the press and audience surveys indicate that subtitles can become a common part of the current viewing experience.¹

Video soundtracks have evolved over the years. Speech that was once clear in the sitcom and drama programs of the analog 60s, 70s, and 80s is now mixed with the modern cinematic techniques possible in today's digital world. Expanded dynamic range has widened the mix engineer's audio palette, but storyline clarity may suffer if recordings and range are not managed correctly for listeners and their environments.

Intelligible dialogue is necessary for audience engagement and enjoyment. When listeners struggle to understand what's being said, they may become annoyed and disengage from the content. Adding to the trouble, the publications in [Annex C](#) indicate that while subtitles are needed for various reasons, there is a consequence: the subtitles detract from the quality of the visual images and engagement with the story.

By considering this report's findings, content producers may choose to evaluate and adjust their production and post-production sound techniques, working to ensure the narratives of their stories are heard clearly. Additionally, the actions needed to improve listening conditions at home are described.

Although much of this paper is written from a video-centric (sound for picture) point of view, most of the information and recommendations (key points) are applicable to audio-only production and distribution for radio broadcasts or podcast programming as well.

3 Human Hearing and Factors Affecting Speech Intelligibility

Understanding how humans hear and understand dialogue is crucial when developing practices to improve intelligibility.

The following is provided for those wanting a summary version of Section 3:

- Intelligibility is not binary. It is a continuum between effortless understanding and total unintelligibility regardless of listening effort. In between are scenarios requiring progressively more listening effort, ranging from situations where all words can be understood to those where some words are missed, no matter how attentively one listens. As effort increases, so does listener fatigue and irritation.
- In the middle of the intelligibility continuum, where understanding is not effortless, listeners use experience and context to help them fill in the gaps. For example, unaccented dialogue in the listener's primary language is easier to understand than dialogue with unfamiliar accents.
- Comprehension depends on a listener's cognitive and hearing ability. Young children and the elderly generally do not understand as well as adults with normal hearing acuity. Because of this, creators with normal hearing can fool themselves into thinking that their productions are intelligible to all audience members.

¹ Articles 1 to 7 in Annex C indicate that a growing number of viewers are using subtitles, ranging between 20 and 50 percent. Most of these reports indicate other causes that are not the fault of audio engineering, such as mumbling and viewers' desire to follow dialogue while they multitask with smart phones, family members, etc.

- In the context of dialogue intelligibility, Masking refers to competing sounds such as music and effects “drowning out” dialogue when they are of comparable levels. This is frequency-dependent: competing sounds in the same frequency range as a given dialogue element are most able to mask it. Masking decreases as the frequency separation between the competing sound and the dialogue increases. This phenomenon is asymmetrical with respect to frequency: with typical program material, a commonly troublesome scenario is heavy bass masking higher frequency dialogue elements, so it is important for creators to monitor on systems with robust bass response typical of worst-case consumer playback systems.
- Depending on its decay time, reverberation can be a masker that reduces intelligibility. Such reverberation may be part of the content and/or added by playback room acoustics.
- Low listening levels can compromise intelligibility by allowing quieter speech elements to be masked by environmental noise and/or to fall below a given listener’s hearing threshold. Compared to consumers, content creators typically listen at higher sound pressure level (SPL) in quieter spaces with less reverberation. This can lead to overly optimistic assessments of intelligibility.

These points and others are elaborated in the full section that follows.

3.1 Bottom-Up and Top-Down Processing

Speech understanding involves both Bottom-Up Processing and Top-Down Processing. Bottom-up processing occurs subconsciously and effortlessly. By contrast, top-down processing requires conscious effort.

Speech understanding begins by analyzing the acoustic building blocks (‘speech cues’) that make up phonemes, which in turn form words and sentences. Examples of speech cues are voice onset times, pitch trajectories, and the spectral envelope of sounds. Decoding speech cues is an input-signal driven, “*bottom-up*” process. This bottom-up analysis is influenced and augmented by *top-down processing*: what the listener expects to hear, e.g., because of grammatical constraints and semantic context. Understanding and comprehending speech involves both processes. The bottom-up process is based on neural structures that are formed in infancy, is engaged in involuntarily, and requires minimal mental effort. The top-down process, on the other hand, requires concentration and mental effort. When the acoustic speech cues are degraded or partially missing, e.g., because the speech is partially masked by background sounds or because the speaker mumbles, the degraded speech can often still be understood. However, to understand in these conditions, listeners must rely more on the top-down process to compensate for the impaired bottom-up process. Heavy reliance on top-down processing leads to listening fatigue. It also compromises the ability to process what was heard (e.g., to comprehend and remember) or to perform concurrent tasks (e.g., driving) because these tasks share the same limited mental resources.

3.2 Intelligibility versus Listening Effort

In technical literature, the term *intelligibility* is reserved for the proportion of an unknown, meaningful message that a listener can understand. An example is the fraction of words a listener can correctly repeat when hearing a sentence. Possible outcomes range from 0% (nothing understood) to 100%

(perfect intelligibility). Intelligibility is objectively quantifiable in a listening test, which says how well the message was received but does not say whether it was easy or hard. That aspect is captured by the measure *listening effort*,² which is a subjective rating of the mental work needed to understand the message. As the speech signal becomes more “clear,” listening effort decreases because the input to the bottom-up process improves, reducing the need for effortful top-down processing. As listening conditions improve, listening effort continues to decrease even though perfect intelligibility has been achieved. Although it is useful to distinguish between listening effort and intelligibility as defined above, outside the field of hearing science the distinction between the two is often blurred and the term *intelligibility* is used more loosely to refer to the ease and ability to understand. This document will use the loose definition.

3.3 Factors Impacting Intelligibility

The factors affecting intelligibility can be grouped into those describing (1) the speaker and the message, (2) degradations of the speech signal during the acoustic and electronic transmission, and (3) the listeners themselves.

Impairments in one or only a few of these dimensions are typically tolerable because speech is a redundant signal. However, as more and more impairments accumulate, listening effort increases and finally intelligibility starts to suffer.

3.3.1 Speaker and Message

- The speaker’s delivery affects how well the message is understood. Fast, casual speech (like everyday conversation) may be harder to follow than clear, deliberate speech; not just because of the faster rate but also because the sounds are not as distinct. Similarly, unfamiliar accents or dialects can increase listening effort.
- The amount of predictability in the message affects its intelligibility: familiarity with a script will require less listening effort.
- Speech is produced in part by moving one's lips. Some speech sounds are associated with clearly recognizable lip movements. When listeners can see a speaker’s lips synchronized to the audio, listeners combine the visual cues and the acoustic cues to obtain better bottom-up input.³

3.3.2 Degradation of the Speech Signal Through Acoustic and Electronic Transmission

As discussed earlier, the speech signal contains acoustic features that drive the bottom-up process of speech understanding. When listeners cannot hear a portion of these features, intelligibility decreases and listening effort goes up. Speech features may become inaudible because they are *masked* by competing sounds or are too weak to reach the hearing threshold. Masking occurs when the neural activation caused by a target signal (speech in this case) is weaker than the activation from a competing sound (masker).

² McGarrigle et al. “Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group ‘white paper’,” *Int J Audiol.*, (2014), <https://doi.org/10.3109/14992027.2014.890296>.

³ https://en.wikipedia.org/wiki/McGurk_effect

Speech is a dynamic signal, and its energy fluctuates across time and frequency at a rate roughly equal to the rate at which syllables are produced. This means that when speech is heard in the presence of a competing, equally strong *steady* sound, at any moment only a fraction of the speech will be masked. The proportion of the speech signal that is masked will depend on the speech-to-masker ratio and on the frequency spectrum of both the masker and the speech.

A very simple, time-proven model⁴ predicts intelligibility by quantifying the proportion of the speech that is being masked. In the following text we use that model to illustrate some of the key factors that affect intelligibility during sound capture, transmission and reproduction.

3.3.3 Ratio Between Speech Level and Level of Interfering Sounds

According to the model, when speech is at a constant level, its short-term energy fluctuates in a range of roughly ± 15 dB. When the speech level is at least 15 dB above the level of a steady masker (in the content or created by the listening environment) all speech cues are audible. In this case the bottom-up process is unimpaired, and intelligibility is optimal. When the speech level drops relative to the masker level, progressively more speech segments are masked until, at a speech to masker ratio of -15 dB, all speech cues are masked, and speech intelligibility is zero.

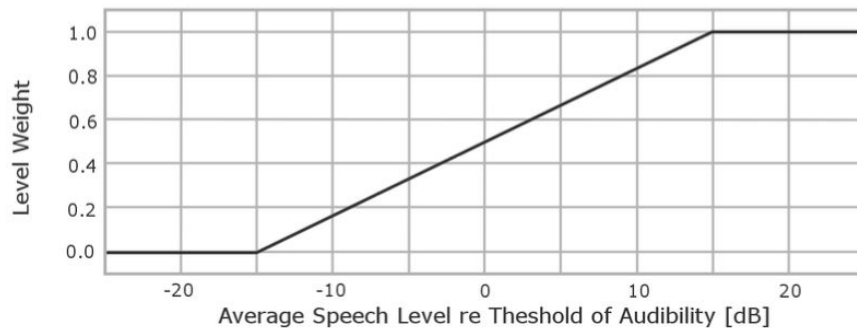


Figure 2. The proportion of speech cues available to the listener (Level Weight) as a function of speech level.⁵ This relationship is described by the Speech Intelligibility Index.⁶

It is often true that the long-term spectrum of the masker is different from that of the speech. In this case, the speech to masker ratio varies with frequency and the above analysis is done in frequency bands (e.g., in $\frac{1}{3}$ octave bands). The weighted sum of the contributions in the bands then determines intelligibility. Greater weight is given to bands at mid frequencies as these are more important for speech understanding than low and high frequencies.

⁴ ANSI/ASA S3.5-1997 (R2020) "Methods for Calculation of the Speech Intelligibility Index"

⁵ Müsch, Hannes, "Aging and Sound Perception: Desirable Characteristics of Entertainment Audio for the Elderly," *Audio Engineering Society Convention Paper 7627*, (2008). Available: <https://aes2.org/publications/elibrary-page/?id=14778>

⁶ ANSI/ASA S3.5-1997 (R2020), "Methods for Calculation of the Speech Intelligibility Index"

A competing sound that fluctuates across time and frequency⁷ masks less effectively than a steady sound of the same level. This means that background music will impair intelligibility less than a steady masking noise with the same level and spectrum as the music. However, interfering speech, which is also a fluctuating masker, impairs intelligibility more than a steady masker because of additional effects related to the listener's inability to distinguish between the speakers.

Masking is determined by the relative strength of the neural activation triggered by speech and masker, which is not necessarily the same as the speech to masker ratio.⁸ This is because a narrowband signal or signal component activates not just the neurons at the signal frequency but also neurons responsible for processing higher frequencies. Because of this *spread of masking*, strong low-frequency signals can mask signals at higher frequencies and have an outsized impact on intelligibility.

Reverberation consists of early reflections (up to 20-50 ms), and late reflections: the reverb's decay. Early reflections can combine constructively with the speech signal and can enhance intelligibility. Conversely, late reflections smear the speech envelope, obliterating speech features. The late reflections create a masker that partially masks the speech that created them. Typically, intelligibility decreases as the reverberation time and the resulting sound energy increases. This applies to both reverb in the content and created by the playback environment.

When talker and masker are spatially separated, intelligibility is typically better than when they are co-located. Listeners use subtle differences between the signals in the left and right ears (binaural cues) to deconstruct the sound scene. This aids in guiding attention and allows listeners to achieve a perceptual speech-to-background ratio that is better than it would have been had there been no spatial separation between the signals (binaural unmasking). Reducing channel count (downmixing) and excessively low bit rate spatial coding (such as parametric stereo or mid-side coding) may reduce or eliminate a listener's ability to benefit from binaural unmasking⁹.

3.3.4 The Listener

The same speech signal can be experienced very differently by different listeners (especially those with a history of noise exposure and older listeners); speech that is easily understood by one listener may be challenging to understand for another. The following are examples of listener characteristics that affect a listener's ability to understand speech:

- Hearing loss, as measured by a traditional pure-tone hearing test ([Audiogram](#)), quantifies the ability to hear quiet sounds, often predominantly at high frequencies. This aspect of hearing loss describes the proverbial inability to hear crickets. To a listener with this type of loss, the bottom part of the speech dynamic range is inaudible. This is like the effect of a background sound (masker) described above, except that this masker is internal to the listener. Listeners with this type of hearing loss will benefit from personalized, frequency- and level-dependent amplification of the speech signal (e.g., hearing aids and assisted listening devices) to overcome that internal masking.

⁷ Festen, J. M., and Plomp, R., "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.* 88, 1725–1736 (1990)

⁸ Moore, B.C.J., "An Introduction to the Psychology of Hearing", 6th Ed., Brill (2012)

⁹ Moore, B.C.J., op. cit.

- The inability to hear quiet sounds is not the only effect of hearing loss. Many listeners with hearing loss are less efficient at extracting speech from background than listeners with normal hearing. This effect is not captured by the Audiogram and can even be present in listeners with normal Audiograms. These listeners require a higher speech to background ratio to understand speech and often have more trouble than others to understand in a noisy place like a restaurant. These listeners will also struggle with mix ratios and noise in the listening environment that does not bother normal-hearing listeners.
- Listeners may have diminished binaural hearing, making the spatial separation described in [3.3.3](#) ineffective.
- The ability to extract speech from a background develops with age, and children require a higher speech to background ratio than adults. Adults listening to a language other than their native tongue also require a higher speech to background ratio than native listeners.

Given the diverse audience and listening environments that media serves, a single intelligibility measurement is insufficient to characterize a program's clarity accurately. To assess whether dialogue is 'clear,' we must consider the target audience's hearing health, age, language abilities, and listening environment. Only by analyzing these factors can we determine if the entire audience is being served appropriately. Such an assessment likely requires a model-based approach, incorporating computational models of speech intelligibility, listening effort, and audience demographics.¹⁰

4 Production and Post-Production

4.1 The Current State of Production Mixing, Monitoring, and Post-Production

Nearfield mixing for home formats has evolved over the last two decades in ways that can have a negative impact on dialogue intelligibility:

- Acting styles have become more subdued and dialogue performances have become quieter.
- Increased location shooting and faster production schedules have resulted in recordings with a higher noise floor.
- Mix schedules have typically not increased to accommodate the added complexity.
- Television content has become much more cinematic, increasing soundtrack complexity and dynamic range.
- Most contemporary cinema soundtracks do not translate well to television; they are mixed at a higher monitor level and typically employ a wider dynamic range. These mixes are too dynamic for comfortable home listening and require re-mixing.
- Many titles primarily intended for home viewing are instead treated like theatrical features in post-production, leaving an appropriate at-home mix as an afterthought. To impress clients, these mixes are done in rooms that are inappropriately large, with monitoring systems that don't accurately represent home playback.

¹⁰ A. Master and H. Muesch, "A Model to Predict the Impact of Dialogue Enhancement or Mix Ratio on a Large Audience," *Audio Engineering Society Convention e-Brief 637* (2020). <https://www.aes.org/e-lib/browse.cfm?elib=20923>

- Key creative stakeholders (directors, producers and show runners) do not get the intended impression of the mix if they review it too far from the optimum monitoring position.
- International loudness recommendations vary and are frequently misinterpreted and / or misapplied.
 - In the US the Commercial Advertising Loudness Mitigation Act (CALM) requires broadcast, cable and satellite content to comply with the Advanced Television Systems Committee (ATSC) Recommended Practice A/85.¹¹ However, its specified Anchor Element (dialogue for long-form content) is often incorrectly identified.
 - The European Broadcasting Union (EBU) has a different set of loudness recommendations that frequently identify full-program-mix for the measured loudness value (and not dialogue).¹²

In addition, as of this writing, streaming services are known to have disparate loudness specifications, which impacts dialogue intelligibility. Some anchor loudness to the full program mix and others to dialogue. This makes global delivery more challenging.

Key Points:

- **Typically, content mixed for cinema should be remixed with appropriate dynamic range for home viewing using a suitable monitoring level and distance per the guidance in this document.**
- **Loudness specifications are complex. Content recipients can misunderstand the rules. A dialog between the mix crew and the recipient is encouraged so that creative intent is preserved and distribution requirements are fulfilled.**
- **Recipient specifications should precisely conform to the regional guidelines. E.g., correctly apply dialogue versus full program mix loudness.**
- **Content creators should optimize dialogue intelligibility within the bounds of the recipient's specification.**

4.2 Practices for Improvement: Methods for Ensuring Intelligibility

The following considerations can assist in maximizing intelligibility. *They should not be considered prescriptive*, but any combination of some or all will contribute to a mix that maximizes dialogue clarity.

4.2.1 Intelligibility in Stereo Listening

Most consumers listen to stereo, so stereo should not be subordinate to the immersive and 5.1 deliverables. Typically, no time or budget is allocated to create a dedicated, independent stereo mix, and even if one exists, not all receivers and set-top-boxes (STBs) will present this version to the audience. Therefore, carefully check and give equal priority to the derived stereo downmix during the mix process.

Ensuring intelligibility of the derived stereo downmix can benefit the main immersive and 5.1 mixes. The stereo version will have increased mutual masking of program elements due to the reduced number of channels, so any dialogue intelligibility problems from the main mix are typically exacerbated in the

¹¹ <https://www.atsc.org/atsc-documents/a85-techniques-for-establishing-and-maintaining-audio-loudness-for-digital-television/>

¹² <https://tech.ebu.ch/docs/r/r128.pdf>

derived stereo mix. Optimizing intelligibility of the downmix will almost always translate across all formats without introducing serious compromise or diluting creative intent.

See Downmix in Distribution on page 19 for more detail.

Key Point:

- **Optimizing intelligibility of the derived stereo downmix improves intelligibility for all formats.**

4.2.2 Playback Levels

Variations in dialogue level and intelligibility will be more pronounced at a lower playback sound pressure level (SPL) because quieter passages will be masked by noise in the home environment. Moreover, lowering listening levels greatly decreases perceived bass energy, as illustrated in the equal-loudness contours shown below. Studies¹³ have shown that median home listening level is 58 dB SPL for television (TV) speakers, and 65 dB SPL for home theater/soundbars. Therefore, optimize the mix at a lower listening level than the mix room’s reference SPL, since that is how most of the audience listens. Do this on a speaker system that more closely approximates the home listening experience.

Key Points:

- **Monitor content during mixing and do quality control (QC) emulating consumer conditions, which can include adding background noise (see Section 6.4).**
- **Monitor and check the intelligibility of your mix at consumer levels on small loudspeakers having approximately flat response¹⁴ (see Section 5.2).**

4.2.3 Loudness Measurement and Loudness Ranges

Based on human perception of loudness, the industry-standard method¹⁵ for measuring loudness in digital transmission systems is ITU-R BS.1770. This reports loudness in units of LKFS (Loudness, K-weighted, relative to digital Full-Scale; aka LUFS), for example –24 LKFS. A loudness change of 1 unit corresponds to a gain change of 1 dB.

Careful consideration of loudness measurements and ranges can be informative to help manage dialogue intelligibility. Specifically, the difference between the integrated full Program Loudness and the Integrated Dialogue Loudness, which we will refer to as “Program to Dialogue Loudness Ratio” (PDLR), can be indicative of the overall dynamic range and intelligibility of the content. While this is not tied directly to intelligibility, it may be relevant. A higher value of PDLR may result in a poor experience for

¹³ Eric Benjamin, “Preferred Listening Levels and Acceptance Windows for Dialog Reproduction in the Domestic Environment,” *Audio Engineering Society Convention Paper* 6233 (2004). <http://www.aes.org/e-lib/browse.cfm?elib=12890>

¹⁴ For a detailed discussion of appropriate loudspeaker characteristics, see “Standard Method of Measurement for In-Home Loudspeakers,” (ANSI/CTA-2034-B), and Toole, Olive, and Welti, “Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers, Rooms and Headphones (4th Edition),” Routledge (2025).

¹⁵ ITU-R BS.1770, “Algorithms to measure audio program loudness and true-peak audio level,” <https://www.itu.int/rec/R-REC-BS.1770/en>

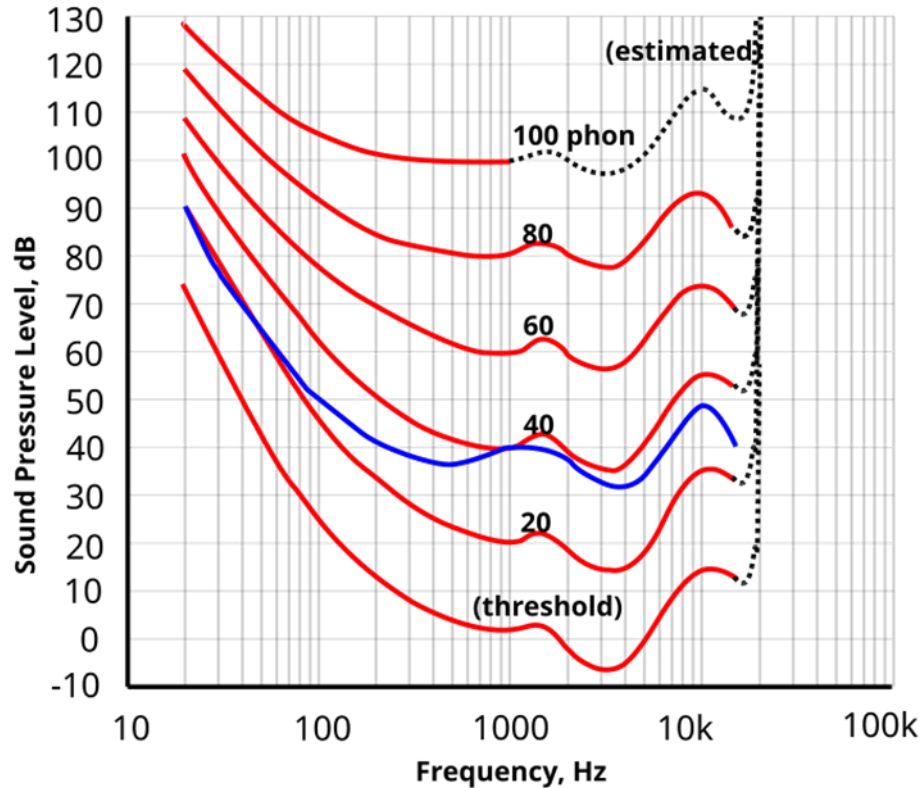


Figure 3. Equal-loudness contours (red) (from ISO 226:2003 revision). Original ISO standard shown (blue) for 40 phons.

the listener, requiring ongoing adjustment of playback level, lowering the volume during louder passages while raising the volume for dialogue sections.

Loudness Range (LRA) is a metric that can provide information about the loudness variation of the full mix or dialogue specifically (when applied to the dialogue track/stem/element). However, it may be misleading.

These factors make it difficult to target a specified LRA when mixing content:

- LRA's level-dependent gating causes it to change in ways that make it difficult for a mixer to perform adjustments that yield predictable results.
- LRA does not take genre into account.
- LRA is independent of dialogue level and can produce misleading values on material that is nevertheless acceptable to the listener.
- Enforcing a maximum LRA diverts available mixing time towards meeting a maximum LRA value instead of maximizing intelligibility.

Key Point:

- **When developing and enforcing content delivery requirements, specifying an LRA target value is misleading, ineffective and discouraged.**

There is a lack of consistency in global industry loudness management practices based on ITU-R BS.1770 that are specified for delivery.¹⁶ As a result, critical details are not uniformly understood or correctly applied. For Long Form Content these conflicting specifications are:

- Dialogue-based loudness
- Full Program Mix Loudness
- Dialogue-based loudness with Maximum Full Program Mix loudness constraint

Dialogue-based measurements have proved to be the most effective means of maintaining dialogue intelligibility and consistency between different programs and channels.

See Loudness Management on page 29 for further discussion.

A single, unified, industry wide loudness delivery specification would be useful in mitigating some aspects of dialogue intelligibility issues outlined in this paper.

In addition to the LKFS and LRA measurement tools mentioned above, listening effort meters that approximate the ability of the human ear to discern dialogue at the phoneme level may also prove useful in evaluating intelligibility (see 5.7.3). As with PDLR and LRA, these developing technologies should be considered informative but not prescriptive.

Key Points:

- **Use dialogue-based anchor measurements to improve/maximize dialogue intelligibility.**
- **A single, unified, industry wide loudness delivery specification would be useful.**

4.2.4 Dialogue Processing in Production and Mixing

The ratio of dialogue to background noise will impact listening effort, so processing tools that reduce noise or enhance dialogue can be beneficial. However, misuse of these tools (e.g., noise reduction, de-reverberation, compression/limiting, expansion) impairs intelligibility and causes listener fatigue.

Tools that apply downward expansion (such as noise reduction and de-reverberation) can render quieter components of words inaudible, degrading intelligibility. While compression can be a useful tool to improve intelligibility of a performance (particularly at low listening levels preferred by many consumers), over-compression or limiting of dialogue can sound unnatural and distracting. Furthermore, if carelessly applied to the full mix, dynamics processing can have a negative effect on dialogue clarity when other, high-energy mix elements compete and cause unintended ducking. Misuse of compression and limiting can be detrimental in the initial stages of production through to the final sound mix of post-production. This can be further exacerbated during distribution where aggressive loudness normalization and compression/limiting are sometimes applied.

¹⁶ ATSC A/85, EBU R128, EBU R128s4, EBU Tech 3343

Additionally, unnecessarily low peak limiting thresholds have proved to be problematic because they may sacrifice necessary impact of the dialogue and compromise audio quality.

Key Point:

- Use dynamics processing cautiously from initial capture through distribution.

4.2.5 Distribution Tools Relevant to Content Creators

Section 5.7.2 discusses intelligibility enhancement features that distributors can offer consumers, including the dialogue object/channel capabilities of next-generation audio codecs (NGAs), and “clean speech” or boosted streams. Because these features may require input from content creators (creating an additional mix with boosted dialogue and/or choosing a Dynamic Range Control profile), it is wise for creators to familiarize themselves with these technologies.

4.2.6 A Culture of Intelligibility

Foster a culture with clients (showrunner, director, producer) where mix intelligibility is prioritized while developing an engaging cinematic experience. Be cautious about compromises to dialogue intelligibility for the sake of creative intent.

Producers should be aware of decisions made on set that compromise dialogue, such as not allocating appropriate time and resources for the sound department to set up and capture performances correctly.

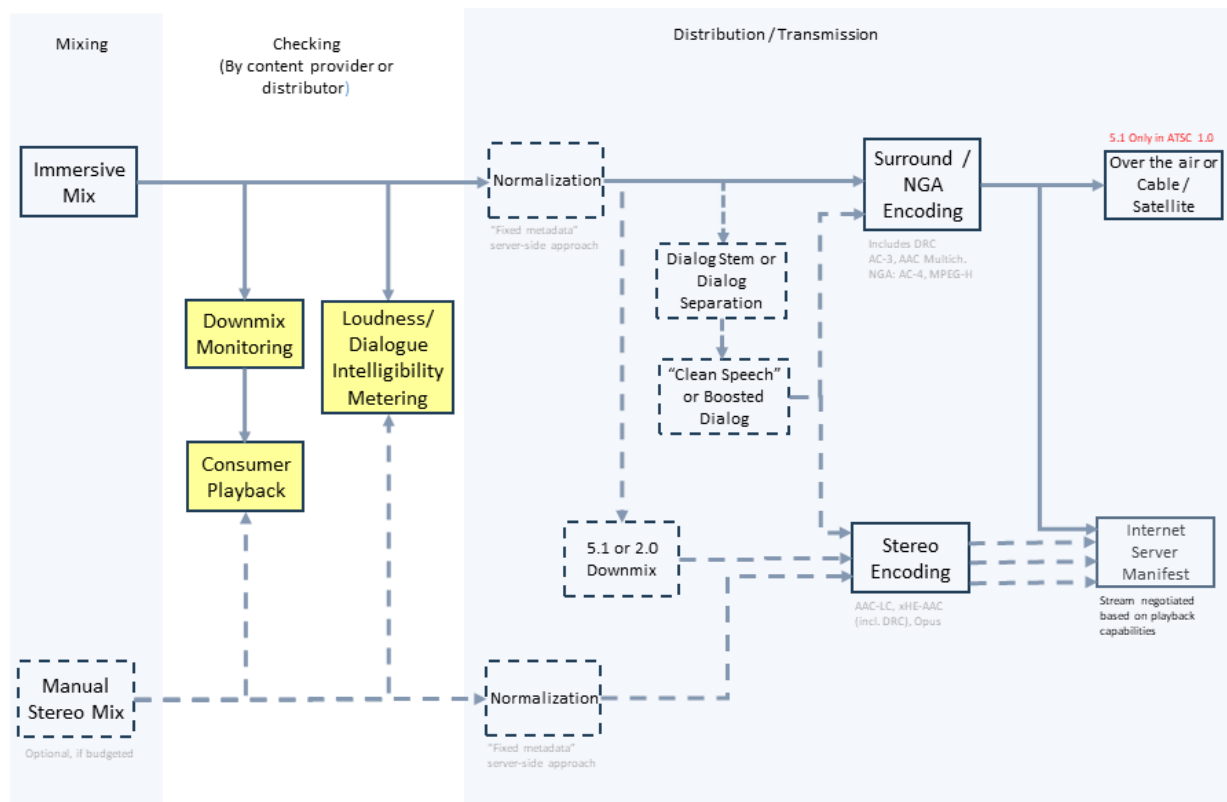


Figure 4. Signal Flow from Mixing through Distribution/Transmission. Suggested practices in yellow.

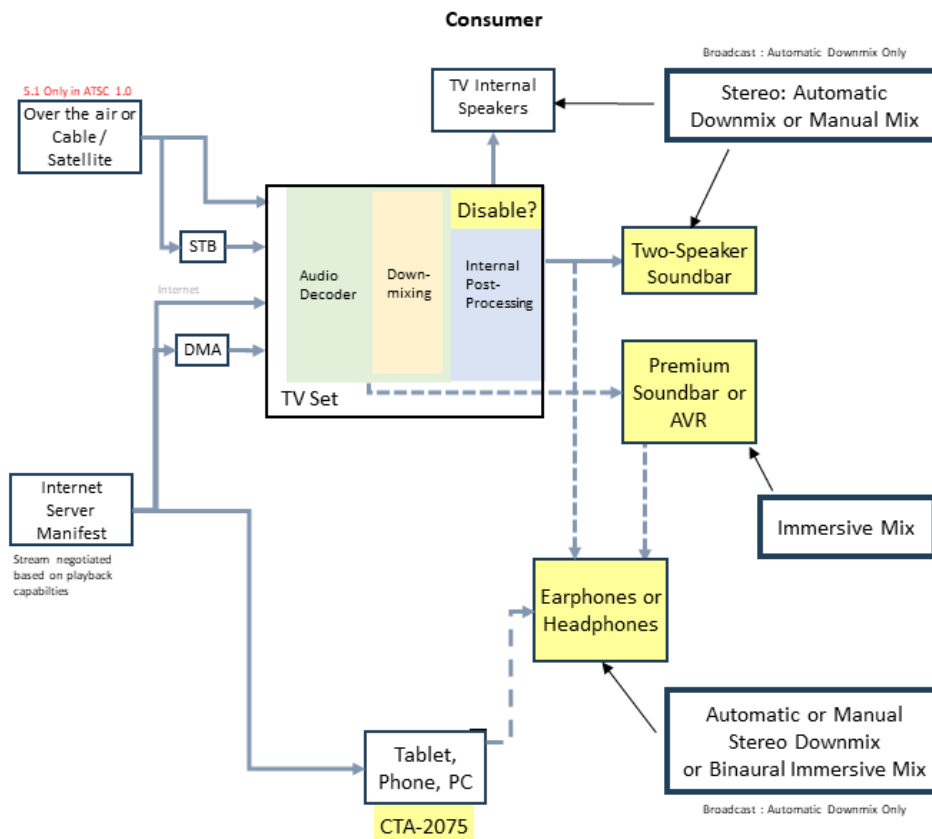


Figure 5. Signal Flow at the Consumer Destination. Suggested practices in yellow.

Casting, direction, and production choices that permit unclearly articulated dialogue will compromise intelligibility.

Automatic Dialogue Replacement (ADR) is a useful process to consider when a production sound recording has provided inadequate intelligibility even after post-processing. It can be indistinguishable from production dialogue when executed properly. Mixing should pay particular attention to dialogue equalization (EQ) to improve unintelligible dialogue while emphasizing natural tonality.

During quality control, it is beneficial for dialogue intelligibility to be assessed by a native speaker unfamiliar with the script. This should be performed with subtitles / captioning turned off.

Section 5.7.3 discusses an emerging class of tools that estimate dialogue intelligibility automatically. These can be useful during a show's quality control process.

Key Points:

- Foster a culture of dialogue intelligibility with the show decision makers.
- During Q/C, it is beneficial for dialogue intelligibility to be assessed by a native speaker unfamiliar with the script.
- Perform dialogue intelligibility Q/C with subtitles / captioning turned off.

5 Distribution

This section explains how dialogue intelligibility can change as finished content proceeds through distribution to the consumer's playback device and listening environment. The signal chain to the device's audio decoder is shown in [Figure 4](#) and [Figure 5](#).

5.1 Distribution Factors

Many distribution channels to the consumer can affect intelligibility. Primary issues are differences in loudness of various content items and sources, and the dialogue signal-to-noise or masking ratio heard in a consumer environment versus a professional environment. This ratio is affected by downmixing and any [Dynamic Range Control](#) (DRC) applied during playback, as well as by acoustic factors in the consumer's environment.

Over-the-air (OTA) broadcasting and over-the-top (OTT) streaming each allow the audience to customize the listening experience. However, streaming offers more options. Even though both allow the audience to adjust the dynamic range and potentially the language of the content, streaming allows the selection of additional options such as dedicated stereo and/or enhanced dialogue mixes. Additionally, the streaming bitrate may adapt to the bandwidth of the delivery path and the playback device's capabilities.

Typical signal paths to the consumer for both broadcast and streaming are shown in [Figure 4](#) and [Figure 5](#), with suggested practices highlighted in yellow.

5.2 The need for Loudness Normalization and DRC

Studies¹⁷ have indicated that consumers listen at lower playback levels than those used by professionals. Moreover, ambient noise and reverberation in the consumer space are likely to be higher than in professional listening^{18,19}. Compared to mix rooms, the consumer space has a lower signal to noise ratio, so to accurately assess dialogue intelligibility in the consumer space, it is important for professionals to take these differences into account when mixing.

Consumers may be fatigued by straining to understand dialogue²⁰ and by the need for frequent volume adjustments if loudness is not managed correctly. Studies have shown that they adjust their volume controls when dialogue loudness is outside their comfort zone, typically +2 to -5 dB with respect to a given consumer's preferred loudness.²¹

Inconsistent loudness management may lead to listening at levels that compromise dialogue intelligibility: Dialogue at the lower end of the comfort zone may be more difficult to understand

¹⁷ Eric Benjamin, "Preferred Listening Levels and Acceptance Windows for Dialog Reproduction in the Domestic Environment," *Audio Engineering Society Convention Paper* 6233 (2004). <http://www.aes.org/e-lib/browse.cfm?elib=12890>

¹⁸ Karl Pearsons, Ricarda Bennett, Sanford Fidell, "Speech Levels in Various Noise Environments," *U.S. Environmental Protection Agency Report* EPA-600/1-77-025 (1977).

¹⁹ J.S. Bradley, "Acoustical Measurements in some Canadian Homes, Institute for Research in Construction," *National Research Council* (1986).

²⁰ See Section [3.2](#).

²¹ Jeffrey Riedmiller, Steve Lyman, Charles Robinson, "Intelligent Program Loudness Measurement and Control: What Satisfies Listeners?" *Audio Engineering Society Convention Paper* 5900 (2003).

because it is more likely to be partially masked by noise and the quietest phonemes may fall below a listener's hearing threshold.

Consequently, material is best presented when it is loudness-normalized. This is achieved by uniformly scaling the loudness of each piece of long form and short form content before presentation. The process makes playback loudness consistent between sources and the content items within them. If normalization is required, it is typically done during ingest/quality control by the distributor. Normalization may also be implemented during playback by a device's audio decoder using Metadata in the codec's bitstream.²² This process is shown in Figure 6.

To accommodate devices with limited playback SPL capability, the normalized output level of the audio decoder can be increased. True home theater systems typically use –31 LKFS as the normalized level, TV sets and soundbars –24 LKFS, and mobile devices –16 LKFS.

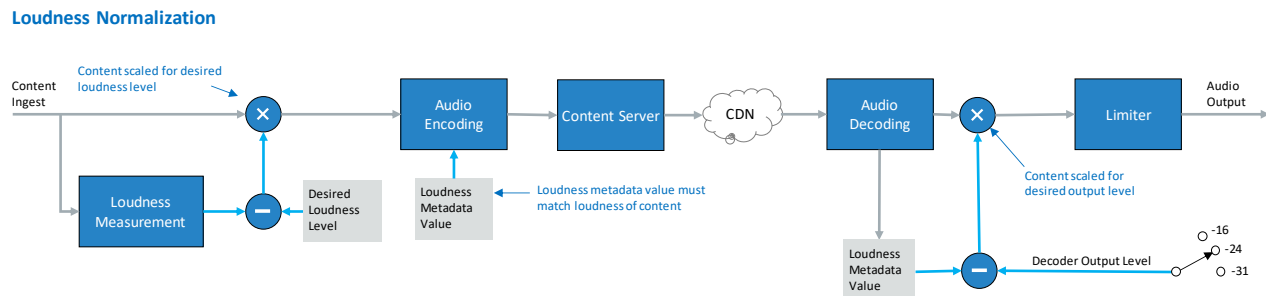


Figure 6. Loudness Normalization Process

Key Points:

- **Correct loudness management minimizes the audience's listening effort and reduces the need to adjust the volume control.**
- **Depending on whether the content is long form or short form, distributors should always ensure that the loudness audio metadata matches either the integrated dialogue loudness or full program mix loudness of the content respectively.**
- **Distributors may additionally wish to normalize content to their chosen loudness target.**

5.3 Loudness Normalization

The transition from analog to digital brought increased audio dynamic range, requiring a new focus on audio loudness to best manage sound for television and streaming.

When creating content, mix engineers should establish an Anchor Element, typically dialogue for long-form content (program material, not ads or promos), and set its average LKFS loudness by fader and meter to the target value specified by the content recipient and/or distributor. Once the LKFS anchor value is established and a reference monitoring SPL in the mix room is set, the remaining music / effects elements of the mix are creatively blended with the anchor by ear. When done in a correctly sized room

²² AES71 and TD1008/AES77 provide normalization workflows and target level recommendations for content, as does EBU R128.

(e.g., 1500-5000 cubic feet) with an appropriate sound pressure monitoring level for –24 LKFS TV mixes (e.g., 76-78 dB SPL)²³, this process helps preserve creative intent and seldom needs correction downstream.

Due to the competitive nature of short-form advertising and promotional content, this content’s anchor is specified as the *full program mix*. Measuring this and not dialogue reduces the ability of rival providers to game or trick the LKFS system by manipulating the dynamics of non-dialogue elements. Accordingly, when a show with a dialogue loudness of –24 LKFS cuts to an ad with –24 LKFS full program mix loudness, the transition is smooth, not requiring any volume adjustments by the listener.

Distributors need to normalize non-compliant content received from creators. Normalization for file-based content uses an LKFS anchor-based measurement. The resultant value is compared to the expected target and on a second pass, the content loudness is uniformly scaled to match the target. There is no impact to quality or dynamic range; the sound is just made quieter or louder.²⁴

For live linear content requiring a loudness adjustment to match the target, an LKFS based automatic loudness control (ALC) processor can be chosen. ALC can be used in real time. However, because it changes dynamic range of the content to achieve a result, creative intent is altered. Even so, these processors can be quite effective when set up by an expert using typical source content. See Section 5.5.

Good loudness management by the content creator and distributor maintains dialogue intelligibility and respects creative intent. The audience has no need to reach for the volume control: the ideal outcome.

Key Point:

- **For normalization of long form content, utilizing a dialogue anchor element will help improve dialogue intelligibility. For short form content, utilize full program mix as the anchor.**

5.4 Dynamic Range Control

Audio codecs²⁵ may include a Dynamic Range Control (DRC) feature that can boost portions of the content so that quiet dialogue is heard above the ambient noise in the user’s environment. It also reduces uncomfortably loud portions that would otherwise require the user to lower the playback volume, possibly compromising dialogue intelligibility.

Codec-based DRC does not alter the encoded audio essence. Gains and control parameters are carried as metadata that can be applied by the user during decoding. DRC in legacy audio codecs typically emulates analog compressors without infinite Look-Ahead. Next-Generation Audio codecs²⁶ expand on this by offering several profiles selectable during playback and may use Infinite Look-Ahead to reduce artifacts. When active, codec-based DRC may improve dialogue intelligibility for users in challenging listening environments, while left inactive for users who wish to hear the unmodified content.

Key Points:

- **Distributors should enable codec-based DRC in the encoder.**

²³ ATSC A/85, Table 10.2

²⁴ Upwards scaling of content may necessitate peak limiting, which may compromise quality.

²⁵ This document primarily addresses video content. Audio-only content may or may not be transmitted with audio codecs or systems that support loudness normalization or DRC.

²⁶ Examples of NGA codecs include MPEG-H 3D Audio, Dolby AC-4, and DTS-UHD.

- Advise users to turn on DRC (if available) when they find it difficult to understand quiet dialogue. Playback devices may have different names for the DRC function (e.g., “night mode”). Users may need to consult their device manual or a search engine to learn how to activate it.

Dynamic Range Control

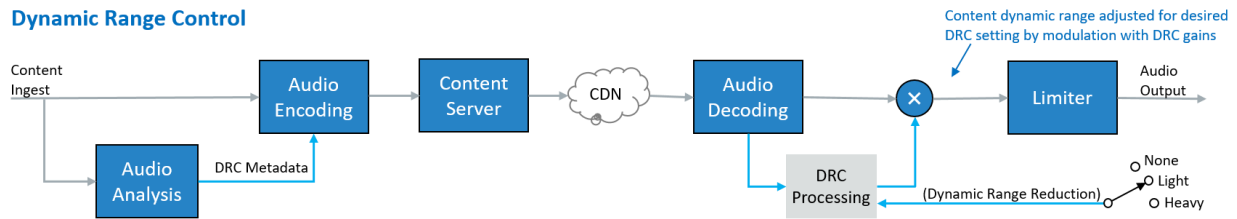


Figure 7. Dynamic Range Control Process

5.5 Real-time Audio Processing

In the absence of upstream loudness management practices that comply with the recommendations earlier in this section, real-time audio processing can improve intelligibility. This processing may be as simple as slow automatic gain control to prevent dialogue from being too faint for easy comprehension. However, most common real-time processors include multiband compression, a process that reduces dynamic range. This is a mainstay in hearing aids and hence scientifically established to improve intelligibility when appropriately implemented. Multiband compression is a method of automatic equalization that can correct muddy-sounding dialogue, preventing masking of consonants essential to intelligibility. It achieves this by increasing the consonants' loudness while simultaneously controlling the loudness of other elements that could otherwise mask them. Level-dependent gating is typically applied to prevent noisy, low-level elements from being excessively amplified and masking desired content. When multiband compression is applied to a multichannel (5.1) or immersive mix, the processor can continuously monitor the short-term ratio of the dialogue loudness and overall program loudness on a band-by-band basis, increasing dialogue loudness if it would otherwise be masked. This method can also be applied to upmixed stereo material if dialogue is properly extracted.

Another technique for automatic equalization, high frequency enhancement, does not reduce dynamic range. Instead, it monitors the ratio of the overall program loudness to the loudness of high frequency content and applies high frequency boost when consonants²⁷ would otherwise be masked by other elements. It can be used independently or in conjunction with multiband compression.

Real-time processing can benefit the intelligibility of content from different eras, where audio may be mono or stereo and was often produced before the advent of modern loudness standards. In addition to enhancing dialogue intelligibility as described above, multiband processing makes such content more mutually consistent and can tame harsh dialogue. This is analogous to colorists' ensuring that video has a consistent “look” between scenes. Another important application of real-time processing for intelligibility is local news programs or other live broadcasts that contain both studio and remote production, which may have unpredictable and widely disparate spectral balance and loudness.

Problems with real-time processing can stem from at least two sources. First is using processors that are not purpose-built for unattended, real-time processing of soundtracks, such as studio and production-

²⁷ See Section 3.3.2 *Degradation of the Speech Signal Through Acoustic and Electronic Transmission*.

targeted compressors without gain-freezing silence gating. Studio compressors are not designed for unattended use with a wide variety of program material, and can cause noise and background sounds to be pumped up obtrusively and unnaturally. Second is incorrect setup of purpose-built processors, particularly the use of excessive amounts of fast multiband compression combined with little or no silence gating. This flattens Microdynamics, resulting in an unnatural, fatiguing, “busy” sound that can compromise intelligibility by amplifying noise and non-speech elements so that they mask speech elements.

Purpose-built processors typically incorporate at least the following chain: gated automatic gain control (AGC), equalization (sometimes including dynamic equalization like high frequency enhancement), multiband compression, loudness control, and peak limiting. Experience indicates that a properly designed AGC should do most of the work of adjusting levels because it operates slowly and does not damage microdynamics. Typically, the AGC operates at a nominal gain reduction of 10 dB and is “window-gated”: its release is very slow when its input level is within a user-specified range (typically ± 3 to ± 6 dB) centered on the current amount of gain reduction.

Multiband compression gain reduction should be constrained to 3-4 dB, as should dynamic loudness control; the AGC has already done most of the work. If dynamic equalization is available, it should be fully leveraged to correct faulty, intelligibility-damaging spectral balances automatically; unlike multiband compression, this preserves dynamic range.

With the typical distribution loudness of modern broadcasts and streaming services, peak levels seldom if ever reach the peak level constraint of the transmission channel. Therefore, the peak limiter operates rarely, if at all. Frequent peak limiter gain reduction implies incorrect setup: either the processor’s output loudness is too high or the limiter threshold is too low so it is not using the full dynamic range of the transmission channel.

Modern purpose-built real-time processors use digital signal processing and have factory presets complying with the operational criteria described above.

Key Points:

- **Seek expert assistance from the processor manufacturer**
- **Use the processor’s guided setup if available**
- **Choose a factory preset appropriate for the content**
- **Use the processor’s input drive/reference level control to produce the desired amount of AGC gain reduction (typically 10 dB)**
- **Set the processor’s target output loudness to the specified distribution value (typically –24 LKFS)**
- **Avoid the temptation to fine-tune the processor’s controls without expert guidance.**

5.6 Downmix in Distribution

5.6.1 The need for downmixing

5.1 surround or immersive content must be downmixed to stereo for playback on most consumer devices such as TV sets that have stereo speakers. This can happen in several ways:

- Creation of a stereo mix by the content provider, which is sent as a separate stream to the consumer. This is the preferred method if available. See Section [4.2.1 \(Intelligibility in Stereo Listening\)](#).
- Automatic downmixing done in the audio decoder of the playback device. In most broadcasting, this is the only method possible.
- Delivery of a supplemental stereo downmix with boosted dialogue which is sent as a separate stream to consumers with impaired hearing or difficult listening environments.

Downmixing may affect intelligibility due to auditory masking as explained in Section [3.3.3 \(Ratio Between Speech Level and Level of Interfering Sounds\)](#). One form of masking occurs because all the content comes from the stereo speakers. This prevents binaural separation of the surround or overhead content by the listener. Another is the addition of surround or overhead components to the stereo channels, which may increase the level of other program elements such as music or sound effects relative to the dialogue. Of course, legacy or secondary content produced only in stereo will not incur any additional masking.

In the case of automatic downmixing, the surround channels are mixed into the corresponding front channels, and the center channel is mixed equally into the left and right channels. This is normally done with -3 dB gain for both cases²⁸. It is possible, though not usual, to specify different downmix gains in the encoded bitstream that will be applied in the decoder. For example, some distributors may prefer -4.5 or -6 dB for surround gains.

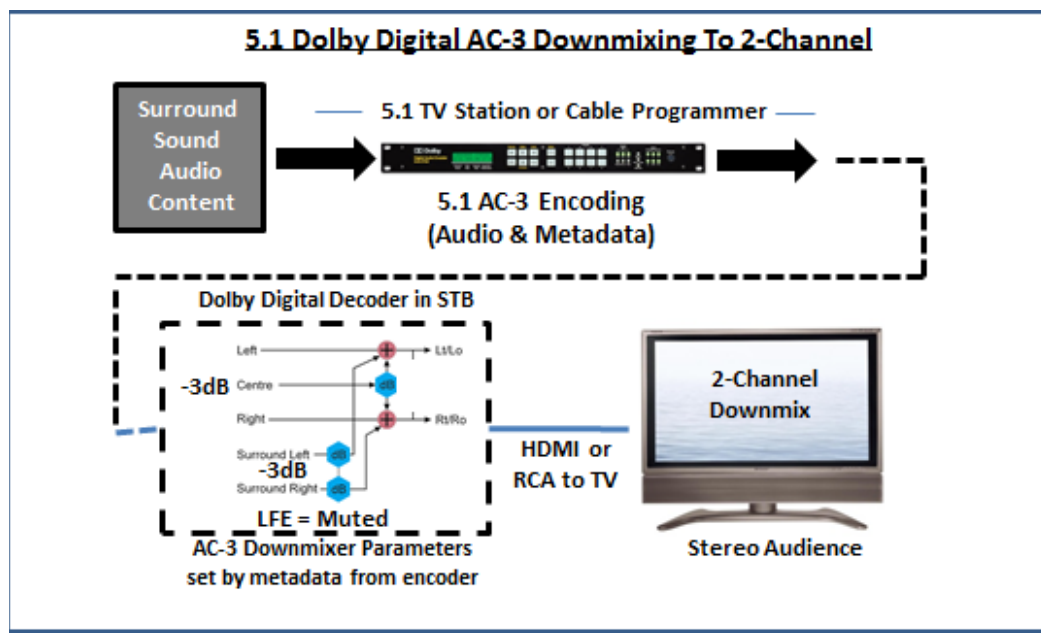


Figure 8. AC-3 Downmixing for Stereo Playback

²⁸ The LFE channel is typically omitted from automatic downmixes. For this reason, it is recommended to avoid including essential content only in the LFE channel.

See AES Technical Council, "Multichannel surround sound systems and operations," *AES Technical Document*, Document AESTD1001.1.01-10 (2001). Available from: <http://www.aes.org/technical/documents/AESTD1001.pdf>, and Tomlinson Holman, "Surround Sound Up and Running, 2nd Ed.," pp. 53–64, *Focal Press* (2017).

5.6.2 Over-the-Top Television Soundtrack to device matching

Recently, automated production systems have been developed that can create many mixes simultaneously for different delivery formats, dialogue languages, and downmix requirements.²⁹ Some distributors create separate boosted or clean dialogue streams that users can select to improve dialogue intelligibility.

Downmixing is a concern not only for TVs but also for 2.0 soundbars. Soundbars with a dedicated center channel speaker will not exhibit downmix masking.³⁰ Some small playback devices such as phones present audio in mono. Therefore, the intelligibility of the mono downmix should be checked and verified.

Key Points:

- **Content providers should check dialogue intelligibility of the program stereo downmix as explained in section 4.2.1. 90% of consumers³¹ hear stereo on their TV sets.**
- **Automatic downmixes should be monitored using the downmix gains that will be used by the distributor.**
- **Content providers should consider auditioning the mono downmix for content streamed to mobile devices.**
- **Distributors should always consider providing a dedicated stereo version of the soundtrack if available from the content provider.**

5.6.3 Metadata Control of the Stereo Downmix Method Used by Decoders

Automatic downmixing may be controlled to produce either a Lt/Rt or a Lo/Ro³² downmix during audio decoding. The Lo/Ro downmix is preferred for intelligibility.

Key Point:

- **Distributors should set encoder metadata for Lo/Ro downmix without a 90° phase shift.**

5.7 Emerging Options

5.7.1 ANSI/CTA-2075-Compliant DRC and Output Levels

CTA-2075³³ defines decoder target output levels and DRC properties that adapt to the listener's devices and environment. It accounts for:

²⁹ https://movielabs.com/case_study/skywalker-keeps-the-humanity-in-automated-soundtrack-mastering/

³⁰ In a mid-range soundbar, there may be slight acoustic masking from the downmixing of surround or immersive channels into speakers that provide only a front image and do not spread surround or immersive channels to a wide sound image.

³¹ Based on reports of committee members who operate popular video services.

³² **Lt/Rt** is a stereo downmix specified in ATSC A/52, Section 7.8.2, and is intended to be used by a matrix surround decoding process (e.g., K. Gundry, "A New Active Matrix Decoder for Surround Sound," in *Proc. AES 19th Int. Conf.*, June 2001). **Lo/Ro** is a stereo downmix specified in Rec. ITU-R BS.775.

³³ "Loudness Standard for Over-the-Top Television and Online Video Distribution for Mobile and Fixed Devices (ANSI/CTA-2075)," *Consumer Technology Association* (Jan. 2020). <https://shop.cta.tech/products/loudness-standard-for-over-the-top-television-and-online-video-distribution-for-mobile-and-fixed-devices-ansi-cta-2075>

- The available resources of the device to achieve appropriate sound pressure level
- Environmental or physical conditions that affect the way a given loudness level and dynamic range are experienced
- User preferences that change the required loudness and dynamic range

DRC and increased target output levels can help to overcome device limitations and environmental conditions. Implementations currently exist that increase output level and apply DRC for the least capable devices.³⁴ Decoder target output levels and DRC profiles are documented in CTA-2075.

When a content distributor only performs downward or no normalization, some devices may not have enough gain to allow dialogue to be understood. However, implementing CTA-2075 ensures that DRC and target levels are applied complementary to the devices capabilities so it can boost dialogue to intelligible levels.

Key Point:

- **Device manufacturers should consider implementing CTA-2075 in their devices.**

5.7.2 Distribution technologies to improve intelligibility

There are two common methods available to allow consumers to adjust dialogue level with respect to other elements. The first is using the dialogue object/channel capabilities of a Next-Generation Audio (NGA) codec to allow consumers to fine-tune the dialogue level. The second is to provide one or more isolated dialogue_or boosted streams that the consumer can choose.

- The first method requires use of an NGA codec where dialogue can be conveyed as a separate stream that allows consumers to adjust the dialogue level either directly or by presets. This requires dialogue to be available as an independent stem (sub-mix) from production or to be separated automatically during the encoding process.
- The second method requires distributors to prepare an isolated dialogue or boosted dialogue version of the content for use in cases of poor intelligibility or for users with hearing impairment. As above, this may be done with dialogue stems from production or with automatic dialogue separation. The advantage of this method is that distributors have more control over the mix, including processing of the dialogue stem.
- An alternative method is for consumers to adjust the playback level of the center speaker in their audio-video receiver (AVR). This is the least desirable method since the center channel may contain other program elements and dialogue may be present in other channels.

Key Points:

- **For NGA distribution, encode dialogue as a separate object or channel to allow user personalization of dialogue level.**
- **Consider offering additional streams with boosted or isolated dialogue.**

³⁴ For example, <https://developer.android.com/about/versions/15/features>

5.7.3 Algorithms to Estimate Intelligibility

Several tools and metrics have been developed and proposed to identify content sections needing human review for potential intelligibility issues. One example is an automatic speech recognizer that uses a deep neural network to estimate the listening effort needed to correctly perceive individual phonemes.³⁵ Another separates dialogue and then measures the short-term loudness ratio between dialogue and background sound.³⁶ Other methods of estimating intelligibility exist or are under development. Distributors can evaluate the utility of these tools for identifying intelligibility issues.

Once listeners have verified intelligibility issues identified by the tools, distributors have several options to mitigate them. Potentially, captions could be displayed for the identified sections, a boosted dialogue stream could be provided, or the producer could be informed that the content does not meet the distributor's requirements.

Key Point:

- **Evaluate the use of automatic scanning tools to identify content sections needing human review for potential intelligibility issues during ingest / encoding.**

6 Consumer Playback Devices and Listening Environments

This section explains how dialogue intelligibility can be compromised as finished content proceeds through distribution to the consumer's playback device and listening environment. The signal chain to the device's audio decoder is first considered, then the contributions of the playback device itself.

6.1 Televisions

Television set sound quality has become secondary to cost and aesthetics. Sets have evolved from cathode ray tube (CRT) displays with front-firing speakers to thin flat-panel displays with narrow bezels and hidden rear or down firing low-excursion speakers that rely on surface reflections. These low cost, tiny speakers are typically mounted in size-constrained enclosures that sometimes include internal waveguides (ducts).³⁷

Key Point:

- **Design constraints and cost sensitivity typically compromise modern TV set sound quality. This section describes practices that may mitigate the issues.**

Speakers are often controlled by driver circuits that monitor voice coil temperature and/or coil displacement indirectly to maximize output. Speaker signal processing will likely also be used to increase

³⁵ R. Huber, H. Baumgartner, S. Goetze, J. Rennies, "ASR-Based, Single-Ended Modeling of Listening Effort—A Tool for TV Sound Engineers," *Forum Acusticum*, pp. 2441-2445 (2020 December).

³⁶ Matteo Torcoli, Mhd Modar Halimeh, Thomas Leitz, Yannik Grewe, Michael Kratschmer, Bernhard Neugebauer, Adrian Murtaza, Harald Fuchs, Emanuel A. P. Habets, "Speech Loudness in Broadcasting and Streaming," *Audio Engineering Society Convention Paper* 10698 (2024).

³⁷ The group's subject matter expert on TV set design stated that "more money is spent on the TV's label and shipping box than its audio subsystem."

the loudness. This induces dynamic compression or limiting of the playback signal at high playback levels, compromising intelligibility.³⁸

Manufacturers have responded to consumer complaints with post-processing intended to improve dialogue intelligibility. These approaches could involve automatic dialogue separation and boosting as previously described. Equalization of the dialogue frequency range and/or ducking of non-center channels are less sophisticated techniques that are often used. All these techniques have varying degrees of performance, can create artifacts, and risk impacting creative intent. Unfortunately, consumers may not realize that their new TVs and soundbars are defaulted to obscure or unexplained settings or modes such as “AI Sound” or “Sports,” which they may not find or understand. Some enhancements such as pseudo surround sound, bass boost, or image widening/virtualization can negatively impact intelligibility and consumers should consider turning them off.



Figure 9. Speaker system of typical TV

Key Points:

- **Manufacturers should default TV set sound enhancements to off.**
- **Consumers should be able to easily choose sound settings that work for them, ideally accessed by a remote-activated popup menu.**
- **Manufacturers should consider offering on-screen help, such as tooltips, to explain the best settings to improve intelligibility.**

Some TV sets include room equalization guided by an automated measurement, as well as settings for wall vs. table use that can help flatten the frequency response in the critical voice range to avoid compromising intelligibility. The frequency response of TVs varies considerably, as shown in [Figure 10](#), for

³⁸ A typical loudness target for a TV is “80-85 dBA SPL at listening position.” This may require feedback-based driving and compression to achieve. A committee member reported compression typically begins at the 50% volume setting.

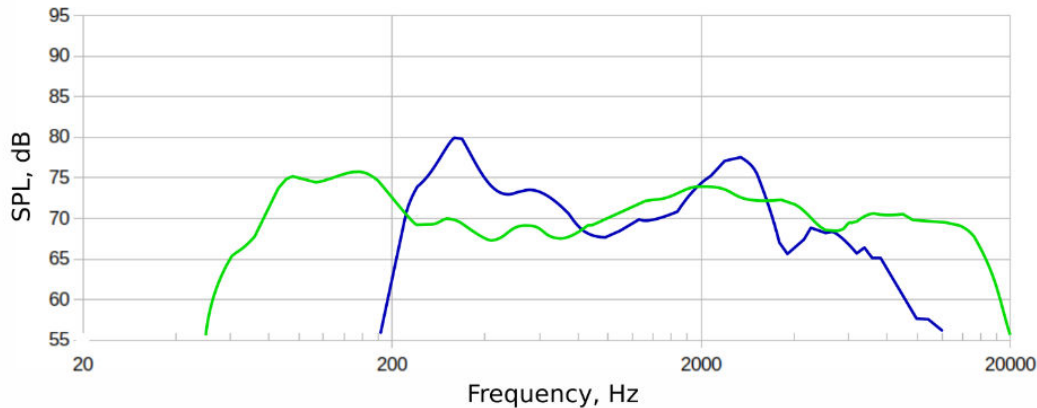


Figure 10. Acoustic frequency response of the best-sounding 2023 TV (green) and representative poor-sounding TV (blue) from consumer test website

the best and worst TVs tested recently by a consumer website.³⁹ Uniform reproduction of the 100-6000 Hz range is important for intelligibility. Compounding the problem, many TVs may not offer enough loudness or fidelity for intelligibility.

Key Points:

- **Many modern flat-screen television designs compromise dialogue intelligibility, reducing audience engagement. Manufacturers should consider prioritizing audio performance, ensuring televisions provide at least 75 dBA SPL at 3m from 300-6000 Hz without compression or limiting, and with inaudible nonlinear distortion.⁴⁰**
- **Encourage consumers to use the room measurement/correction tool of their TV, if available.**

6.2 Upgraded Audio Systems

The best TV dialogue intelligibility is obtained with external speakers, connected to an audio-video receiver (AVR), in an appropriate placement for the sound format and content consumed. Some premium TVs and soundbars can be networked with satellite speakers to work together, potentially offering a convenient alternative to classic AVR systems. AVR systems are beyond the typical budget of mainstream consumers, but a close level of performance can be obtained with a premium soundbar with more than three channels. Next would be a lower-cost soundbar with at least a Left–Center–Right speaker arrangement. Even a stereo soundbar may provide better intelligibility than a typical TV set.

Key Points:

- **To obtain the best dialogue intelligibility, consumers should upgrade from the TV's internal sound to a correctly set up external solution described above.**
- **For stereo playback devices, consumers should consider selecting a manual stereo mix if available in the application player. Note that stereo streams may be a manual or automatic downmix and thus the intelligibility benefit may vary.**

³⁹ <https://www.rtings.com/>

⁴⁰ Based on a consumer listening level of 60 dB A SPL, a peak to rms ratio of 15 dB for speech, and the importance function of the Speech Intelligibility Index.

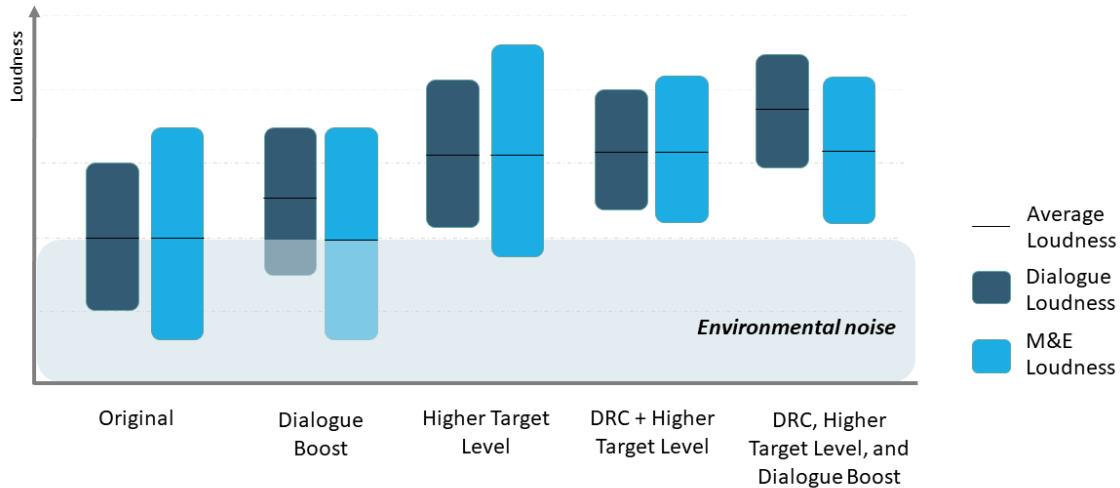


Figure 11. Increasing Dialogue Intelligibility with DRC and Dialogue Boosting.

- **If intelligibility is impaired, consumers should consider Next-Generation Audio codec dialogue enhancement or a stream with boosted or clean dialogue, if available.**

6.3 Device Options for Improving Intelligibility

Intelligibility problems may be alleviated with several strategies that exploit device options.

6.3.1 DRC

Figure 11 shows the effect of several playback options that may be available to the consumer to increase dialogue intelligibility when environmental noise is high. To correct an inadequate volume control range, consumers may select a louder output level, if available (see section 5.3 [Loudness Normalization](#)). Audio decoders also may allow [Dynamic Range Control](#) (DRC, section 5.4) to be applied, which may increase the level of quiet dialogue.

6.3.2 A Distributor's Alternate Stream or NGA CODEC

DRC does not change the balance between dialogue and simultaneously competing elements in the mix, so it cannot reduce masking of dialogue by these elements. Consumers may opt for two different solutions when available: an alternate stream with boosted dialogue or an NGA codec setting (see section 5.7.2) that boosts dialogue.

6.3.3 Post Processing

Also, a device's post-processing can be activated which may enhance and/or boost the dialogue. DRC and alternate stream options may be used simultaneously for improved intelligibility.

6.3.4 Device Configuration Options

Intelligibility can be affected by the complex setup options presented by multiple consumer devices in the playback chain. As the DRC metadata in the audio bitstream is optionally applied during decoding, the DRC and processing options available to the consumer may depend on where the audio is actually decoded.

There are numerous audio signal paths used by consumers. Common examples include:

- The TV decodes the audio from a broadcast tuner or streaming application and sends it to its speakers. If necessary, the TV will perform a downmix to conform to its speaker layout.
- The TV decodes the audio from a broadcast tuner or streaming application and sends it to an AVR or soundbar via S/PDIF or HDMI.
- The TV passes un-decoded audio via S/PDIF or HDMI to be decoded by an AVR or soundbar. This is commonly called *bitstream pass-through*.
- A set top box or digital media adapter decodes the audio and sends it to an AVR or soundbar. The TV only performs picture playback.
- A set top box or digital media adapter passes the un-decoded audio to an AVR or soundbar. This is commonly called *bitstream pass-through*. The audio may be immersive. The TV only performs picture playback.
- A set top box or digital media adapter decodes the core audio into PCM and sends this and immersive metadata to the AVR or soundbar, which extracts the immersive audio channels. The TV only performs picture playback.

Figure 12 illustrates a simplified version of a signal flow through a typical TV's audio path. In the simplest case, incoming encoded audio is decoded by the TV's audio decoder or selected distributor's app. The decoded audio is mixed with system sounds such as clicks or beeps provided as feedback to the user and sent to the TV's speakers. The consumer may then need to adjust DRC and post-processing settings if available in the TV's menu or in a distributor's app.

When an AVR or soundbar is connected to the TV as shown,⁴¹ the audio may be sent to it as decoded PCM, as the received audio bitstream, or as a re-encoded version of the decoded audio. The TV or distributor's app may give the consumer one or more of these options in its setup menu.

6.3.5 Bitstream Passthrough

Sending the received audio bitstream, termed bitstream passthrough, bypasses decoding in the TV, preserving the DRC metadata (see Section 5.4) for decoding in a downstream device. Bitstream passthrough bypasses any post-processing in the TV.⁴² This is advantageous when the device receiving

⁴¹ If devices support the required audio/video formats via HDMI ARC/eARC, an alternative configuration is to connect the AVR or soundbar to signal sources ahead of the TV, avoiding bitstream passthrough. This does require selecting the signal source using the soundbar or AVR's remote control instead of the TVs.

⁴² Post-processing in the playback device may also control dynamic range or boost dialogue. However, this processing does not take advantage of the infinite look-ahead and creative review that DRC and other server-side technologies offer.

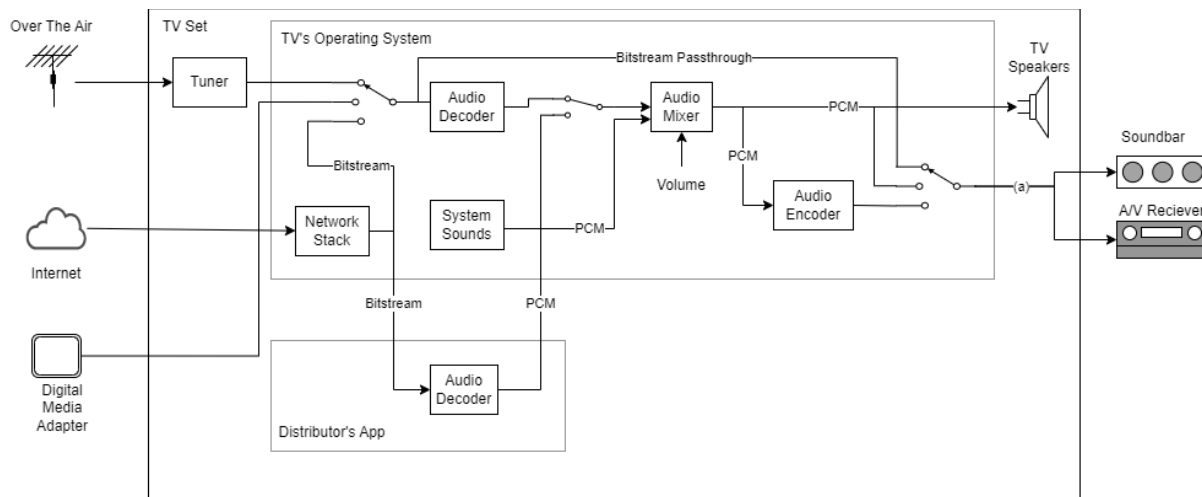


Figure 12. Audio signal paths from a typical TV's audio decoders to its speakers and a connected soundbar or AVR. Note that signal (a), typically carried over HDMI, may be stereo PCM, an encoded bitstream, or an uncompressed bitstream.

the bitstream (such as an AVR) offers the consumer more options (such as the DRC amount applied) than the source device.

The signal flow of Figure 12 is also representative of Digital Media Adapters or Set-top Boxes, which may offer PCM or bitstream output options as well. Because audio adjustments may potentially be done in all the devices in the signal chain, this further increases the complexity (and number) of user adjustments. This may be confusing since adjustments in the device not performing decoding will have no effect. If decoding is done in a distributor's app, adjustments there will only affect that distributor's content.

Key Points:

- **Manufacturers should make it easy for users to find and activate DRC.**
- **While DRC can help in the case of quiet dialogue, typically it cannot decrease masking by other program elements, so dialogue-boosting options may be needed.**
- **Consumers need to be alerted that there may be multiple output modes and options in their playback system. Experimentation may be needed to find ones that are active in a given playback situation and provide the best intelligibility. Selecting bitstream passthrough to the final playback device may be advantageous.**

6.4 Room Acoustics

Consumer rooms often have longer reverberation decay times than mixing rooms specified in A/85⁴³ or BS.1116.⁴⁴ Studies over decades have shown typical reverberation time (RT_{60}) values in the 400-600 ms

⁴³ Advanced Television Systems Committee, "A/85: Techniques for Establishing and Maintaining Audio Loudness for Digital Television," <https://www.atsc.org/atsc-documents/a85-techniques-for-establishing-and-maintaining-audio-loudness-for-digital-television/>.

⁴⁴ International Telecommunications Union. ITU-R BS.1116-3, "Methods for the subjective assessment of small impairments in audio systems," <https://www.itu.int/rec/R-REC-BS.1116>.

range for living rooms,^{45,46} and there is anecdotal evidence this is increasing with contemporary décor and architectural trends.⁴⁷ Thus, professionals will be listening in a less reverberant room than the consumer, which may lead mix engineers to perceive greater dialogue intelligibility than consumers. While reverberation is likely a secondary factor, it may be wise to add diffusion to the voice band in dedicated quality control rooms to increase RT_{60} to ~500 ms. Consumer rooms may benefit from added rugs, drapes and other absorption.

Key Point:

- **The consumer environment impacts intelligibility, as it is typically more reverberant than the mix environment.**

6.5 Mobile Devices

Tablets and mobile phones have issues that are an order of magnitude worse than TV sets. Speakers may be only 10-12 mm in diameter and have a 0.5 mm maximum displacement. The frequency response of a typical tablet or phone rolls off below 700 Hz due to acoustic effects.⁴⁸ Active feedback drivers are necessary for tolerable output SPL. Compounding the problem, these devices are often used in environments that have more ambient noise than the home TV room. Speaker playback in this case can require both dynamic range control (DRC) and dialogue boost for successful intelligibility. This may already be performed on the device.

Key Point:

- **Mobile device users listening in noisy environments should try DRC and dialogue boost options when available.**

Earphones and headphones (particularly in-ear or noise cancelling ones) can be superior to mobile device speakers. They reduce the ambient noise, improve maximum SPL, avoid the need for compression or limiting, isolate listening from the environment, and can improve frequency response and reduce distortion. These advantages will benefit dialogue intelligibility. Binaural rendering technology has the potential to greatly improve immersion and envelopment, and can affect intelligibility positively or negatively.

Key Point:

- **Encourage earphone or headphone use in challenging environments.**

6.6 Loudness Management in Playback

Loudness variations across content can negatively affect intelligibility. As discussed in detail above, content creators and distributors should manage loudness according to industry standards so that

⁴⁵ Burgess and Utley, "Reverberation Times in British Living Rooms," *Applied Acoustics* 18 pp. 369-380 (1985).

⁴⁶ Tomlinson Holman and Ryan Green, "First Results from a Large-Scale Measurement Program for Home Theaters," *Audio Engineering Society Convention Paper* 8310 (2010). <https://aes2.org/publications/elibrary-page/?id=15732>.

⁴⁷ Towards more reflective surfaces such as hardwood, tile, and stone, and towards "great room" spaces where kitchen noises are present.

⁴⁸ For a representative mobile phone acoustic frequency response measurement, see Figure 3 in J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," *IEEE International Carnahan Conference on Security Technology* (2011)

consumers do not have to readjust their volume controls between content items. Consistency between program sources will be ensured if playback device manufacturers comply with ANSI/CTA-2075.⁴⁹ Note that some AVRs offer source gain adjustment to help align the loudness of noncompliant sources.

⁴⁹ “Loudness Standard for Over the Top Television and Online Video Distribution for Mobile and Fixed Devices (ANSI/CTA-2075),” *Consumer Technology Association* (January 2020). <https://shop.cta.tech/products/loudness-standard-for-over-the-top-television-and-online-video-distribution-for-mobile-and-fixed-devices-ansi-cta-2075>.

7 Annex A: Terms, Definitions, and Abbreviations

Anchor / Anchor Element: The perceptual loudness reference point or element (typically dialogue) around which other elements are balanced in producing the final mix of the content, or that a reasonable viewer would focus on when setting the volume control. The anchor loudness is used to calculate loudness normalization gain, which is applied uniformly to the content. This process conveys the anchor elements at the listener's preferred loudness while preserving the source's original dynamics. For example: "We normalized this file to -24 LKFS using a dialogue anchor."

Audiogram: A graph that shows a given listener's audible threshold for standardized frequencies as measured by an audiometer. The Y axis represents intensity measured in decibels (dB) and the X axis represents frequency measured in hertz (Hz). The threshold of hearing is plotted relative to a standardized curve that represents 'normal' hearing, in dB (HL) (hearing level).

Automatic Dialogue Replacement (ADR): A post-production process where actors re-record their dialogue in a studio to replace the original audio captured during filming. This is done to improve sound quality, fix mistakes, or change dialogue after the fact. The new recordings are then synchronized with the video footage, matching the original lip movements and emotional intent.

Bottom-Up Processing: Speech understanding that occurs subconsciously and effortlessly. Typically, listeners will subconsciously augment it with Top-Down Processing.

BS.1770 Loudness: An electrical measurement using the ITU-R BS.1770 algorithm. This estimates the perceived loudness of content, referenced to 0 dBFS, in a unit of "LKFS." The acronyms LUFS (Loudness Units with respect to digital Full Scale) and LKFS (K-weighted Loudness with respect to digital Full Scale) are interchangeable and specify identical values. The algorithm consists of a K-weighting filter feeding a level-gated RMS detector followed by a logarithmic converter scaled such that a loudness change of 1 LU corresponds to a level change of 1 dB. The first stage of the filter is a high frequency 4 dB shelving boost that accounts for the acoustical effects of the head. The second stage is a low-frequency roll off based on "B-weighting," which roughly approximates the 70 phons equal-loudness contour (see [Figure 3: ISO 226:2003](#)).

Dialogue: As used in this document, any speech elements within content.

Dialogue intelligibility: How well speech within content can be understood. It is a continuum between effortless understanding and total unintelligibility regardless of listening effort. In technical literature, the term *intelligibility* is reserved for the proportion of an unknown, meaningful message that a listener can understand.

Digital Media Adapter (DMA) (also known as a **Digital Media Receiver** or **Digital Media Player**): A device that allows digital media sources to connect to home media systems like TVs, soundbars and audio-video receivers (AVRs). It enables the transfer and display of digital content, such as photos, music, and videos, over a network, either wired or wireless. The most common DMAs (for example: Roku, Apple TV, and Fire TV) run multiple applications that each support streaming from one content provider.

Distributor: In the context of this paper, the entity in the delivery chain responsible for the normalization of content loudness before release to end users. This entity is typically a network, branded streaming service or on-demand content provider.

Downmixing: The process of reducing the number of audio channels by mixing some of the source content’s channels together. For example, 5.1 content can be downmixed to stereo. Downmixing can impact dialogue intelligibility, as discussed in Section [5.6](#).

Dynamic Range Control (DRC): The process of continually adjusting audio signal level to control the loudness difference between loud and quiet passages. This can help overcome noise in the listening environment, meet the dynamic range capability of the playback equipment, or both. The term “dynamic range control” typically implies that the codec’s encoder generates the gain control signal as metadata and that the player device’s decoder acts on this metadata. While this can also be described as “dynamic range compression,” that term is more commonly used when a single device (like the player) generates the gain control signal from an internal algorithm applied to the incoming audio and does not exploit DRC metadata.

Integrated Dialogue Loudness (aka Speech Loudness): The BS.1770 integrated loudness of dialogue content that is not mixed with other elements such as a music bed. Measurement requires isolating speech segments of content either manually, or by measuring the dialogue stem, or with an automatic speech detection or gating tool.

Interstitial content; Interstitial: Short form content—commercial advertising, commercial, promotional or public service-related material. The typical duration is less than approximately two to three minutes.

Isolated Dialogue (aka Clean Speech): Describes the contents of an audio channel or object containing only dialogue without music or effects.

LKFS: Loudness K-weighted, relative to digital Full Scale. LKFS is synonymous with LUFS.

Look-Ahead: In a dynamics processor such as a compressor or peak limiter, technology that presents the gain-control sidechain with audio that is earlier than the audio applied to the processor’s gain control element. By informing the sidechain about upcoming audio, this allows the sidechain to apply strategies, such as ramping the gain down before a peak is encountered, that minimize audible artifacts introduced by the limiting.

Infinite Look-Ahead: A variation of look-ahead technology used in digital audio file processing, where the entire file is first read and a processing algorithm then uses this global knowledge to compute local gain reduction throughout the file with minimal side-effects.

Long-form Content: “Show or program-related material or essence. The typical duration is greater than approximately two to three minutes.” [per ATSC A/85] Examples include radio shows, concerts, dramas and musical albums.

Loudness: The perceived audio level of a program or other piece of content. This document refers to loudness measurement using the algorithm specified in ITU-R BS.1770.

Loudness Range (LRA): A metric (formally defined in EBU Tech 3342) that can provide information about the loudness variation of the full mix or dialogue specifically (when applied to the dialogue track/stem/element).

Masking: As used in this document, the phenomenon where competing sounds such as music and effects “drown out” dialogue. It is discussed and quantified in Section [3.3.3](#).

Metadata: Data (other than essence) included within content, streaming, or broadcasting used to identify, label and present content. Among many others, metadata can include information such as content loudness, codec in use, bitrate, artist, genre, label, and titles.

Microdynamics: The subtle and intricate volume variations and dynamic contrasts within a small portion of a sound or signal, often within a single measure of music or a specific instrument's performance.

Nearfield Mixing: A monitoring technique where the mix engineer sits close to the monitor speakers so that the direct sound from the speakers is substantially more intense than the mixing room's reverberation at frequencies whose acoustic wavelength is small compared to the room's dimensions. This minimizes the effect of room acoustics on what the mixer hears and can yield an overly optimistic estimate of dialogue intelligibility compared to what the consumer typically hears.

Note that nearfield mixing is not immune to low frequency response peaks caused by room modes and can benefit from low frequency room treatment (such as bass traps) and/or "room equalization."

Next-Generation Audio (NGA) codec: A codec, such as MPEG-H 3D Audio, Dolby AC-4, and DTS-UHD, that, when compared to legacy codecs like Dolby AC-3 or AAC-LC, provide additional features such as support for immersive audio, audio objects (including separated dialogue), and more versatile metadata-driven Dynamic Range Control.

Program Loudness aka Full Program Mix Loudness: The average loudness of all elements of the content, between two points in time, measured using the BS.1770 loudness algorithm with gating.

Program to Dialogue Loudness Ratio (PLDR): The difference between Program Loudness and Integrated Dialogue Loudness.

Set-top Box: A type of Digital Media Adapter. As used currently, the term usually describes a device that accepts an RF or IP signal from a broadcast antenna or cable TV provider, decodes it, and presents it to a TV or audio-video receiver via HDMI.

Speech Loudness: See Integrated Dialogue Loudness.

Top-Down Processing: Listeners using context along with their experience and expectations to help understand speech that cannot be understood by Bottom-Up Processing alone.

8 Annex B: Related Loudness Standards and Recommendations

AES71-2018, "Recommended Practice Loudness Guidelines for Over-the-Top Television and Online Video Distribution," (July 2018). <https://www.aes.org/publications/standards/search.cfm?docID=107>

AES77-2023, "AES Recommended Practice Loudness Guidelines for Internet Audio Streaming and On-Demand Distribution," (July 2023). <https://www.aes.org/publications/standards/search.cfm?docID=114>

ANSI/CTA-2075, "Loudness Standard for Over-the-Top Television and Online Video Distribution for Mobile and Fixed Devices," <https://shop.cta.tech/products/cta-2075>

ATSC A/85, "ATSC Recommended Practice: Techniques for Establishing and Maintaining Audio Loudness for Digital Television," [https://www.https://www.atsc.org/wp-content/uploads/2025/06/A85-2013-with-Corrigendum-No-1.pdf](https://www.atsc.org/wp-content/uploads/2025/06/A85-2013-with-Corrigendum-No-1.pdf)

EBU R 128, "Loudness Normalisation and Permitted Maximum Level of Audio Signals," (November 2023). <https://tech.ebu.ch/docs/r/r128.pdf>

EBU R 128 s1, "Loudness Parameters for Short-Form Content (Adverts, Promos, etc.)," (November 2023). <https://tech.ebu.ch/docs/r/r128s1.pdf>

EBU R 128 s2, "Loudness in Streaming," (November 2023). <https://tech.ebu.ch/docs/r/r128s2.pdf>

EBU R 128 s4, "Loudness Normalisation of Cinematic Content," (November 2023). <https://tech.ebu.ch/docs/r/r128s4.pdf>

EBU - Tech 3341, "Loudness Metering. 'EBU Mode' Metering to supplement EBU R 128 loudness normalization," (November 2023). <https://tech.ebu.ch/docs/tech/tech3341.pdf>

EBU - Tech 3342, "Loudness Range: A measure to supplement EBU R 128 loudness normalisation," <https://tech.ebu.ch/docs/tech/tech3342.pdf>

EBU - Tech 3343, "Guidelines for Production of programmes in accordance with EBU R 128," November 2023, <https://tech.ebu.ch/docs/tech/tech3343.pdf>

EBU - Tech 3344 version 2.1, "Guidelines for distribution and reproduction in accordance with EBU R 128," (July 2016). <https://tech.ebu.ch/docs/tech/tech3344.pdf>

ITU-R BS.1771-1, "Requirements for loudness and true-peak indicating meters," (January 2012). <https://www.itu.int/rec/R-REC-BS.1771-1-201201-I/en>

ITU-R BS.1770-5, "Algorithms to measure audio programme loudness and true-peak audio level," November 2023, <https://www.itu.int/rec/R-REC-BS.1770/en>

9 Annex C: Press Accounts

[1] David Pogue, “The popularity of turning captions ‘on’”, *CBS Sunday Morning*, March 10, 2024, <https://youtu.be/H6u-f69-J-U?si=42pd8BELCFvXVVy->. (CBS’s caption states: “Whether you have lousy TV speakers, are hard of hearing, are distracted by the kids, or are watching a film with actors who mumble, chances are you are using the captions option while watching TV, even if they’re not translating a foreign language. Correspondent David Pogue talks with Oscar-winning sound mixer Tom Fleischman, who explains – despite technological advances in audio – why films may be getting harder to hear...”)

[2] Devin Gordon, “Why Is Everyone Watching TV With the Subtitles On?” *The Atlantic*, June 6, 2023, <https://www.theatlantic.com/ideas/archive/2023/06/watching-movies-tv-with-subtitles/674301/>. (Feature story observing that a significant number of current television viewers are using subtitles [closed captions].)

[3] Jacalyn Wetzel, “Finally, someone explains why we all need subtitles when watching TV,” *Upworthy*, February 1, 2023, <https://www.upworthy.com/subtitles-for-watching-tv-rp5>. This feature story states that it is movie dialogue that is causing viewers to turn on subtitles, not a problem with one’s hearing. Video by Vox Producer Edward Vega is embedded in the story that points to reference [4].

[4] Edward Vega, “Why we all need subtitles now,” *Vox*, (March 2023), <https://www.youtube.com/watch?v=VYJtb2YXae8>. (Entertaining but thorough look at how movie industry voice recording has changed over the years, and now struggles with issues such as background noise effects and actors who are allowed to mumble their lines. The producer points out how modern flat-screen TVs have speakers are inadequate for the demands of theatrical soundtracks and dialogue.)

[5] Nick Pope, “I Couldn’t Hear A Word Of ‘Tenet’ (And I’m Not The Only One)”, undated, <https://www.esquire.com/uk/culture/a33859053/tenet-sound-problems/>. (Commentary that points to numerous viewer complaints who say that popular directors such as *Tenet*’s Christopher Nolan use actor dialogue that is muffled or submerged in noise or music. Nolan is quoted as saying “I actually got calls from other filmmakers who would say ‘I just saw your film, and the dialogue is inaudible.’...but the truth was it was...how we had chosen to mix it.”)

[6] Katie Cunningham, “Mumbling actors, bad speakers or lazy listeners? Why everyone is watching TV with subtitles on,” *The Guardian*, January 27, 2023, <https://www.theguardian.com/tv-and-radio/2023/jan/28/mumbling-actors-bad-speakers-or-lazy-listeners-why-everyone-is-watching-tv-with-subtitles-on>. (News article reports that Netflix finds 40% of its viewers are regularly using subtitles. The article quotes industry sound mixers who assert that the cause is a combination of new mumbling acting styles and more background sounds in the dialogue.)

[7] Nadia Mykhalevych, “Survey: Why America is obsessed with subtitles,” *Preply*, October 13, 2023, <https://preply.com/en/blog/americas-subtitles-use/>. (A survey of more than 1,260 Americans on how and why they use subtitles, which finds 50% of Americans watch with subtitles and 55% say it is harder to hear dialogue on TV than before. Nearly three-quarters of respondents claimed “muddled audio” was to blame, but a fraction prefer to watch at home with low volume or are multitasking.)