

# EBU

OPERATING EUROVISION AND EURORADIO

# TECHNICAL REVIEW

## Speech Intelligibility in TV

June 2022

H. Baumgartner, Fraunhofer IDMT

R. van Everdingen, Delta Sigma Consultancy

B. Schreiner, BR

M. Kahsnitz, RTW

U. Krämer, ARD.ZDF Medienakademie

The main purpose of an EBU Technical Review is to critically examine new technologies or developments in media production or distribution. All Technical Reviews are reviewed by 1 (or more) technical experts at the EBU or externally and by the EBU Technical Editions Manager. Responsibility for the views expressed in this article rests solely with the author(s).

To access the full collection of our Technical Reviews, please see:  
[tech.ebu.ch/publications](http://tech.ebu.ch/publications)

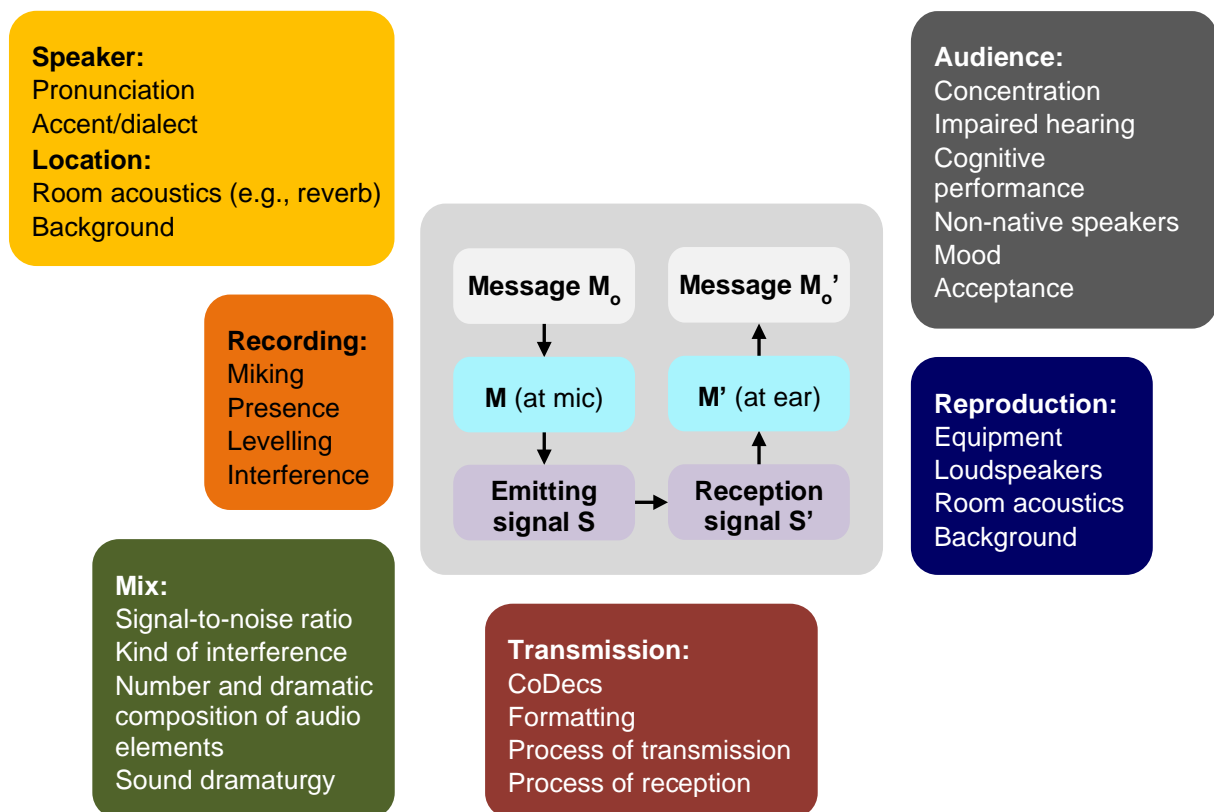
If you are interested in submitting a topic for an EBU Technical Review, please contact: [tech@ebu.ch](mailto:tech@ebu.ch)

## 1. Introduction

This overview originated from the session ‘*Speech Intelligibility in Film and Television*’ and the subsequent roundtable of presenting experts, at the Tonmeistertagung 2016, held in Germany. Six years later, the basic aspects of the discussion are still relevant, which was the reason for this English translation and publication to a wider audience.

‘*Put 70-year-old sound engineers at the mixing desks and they’ll mix the sound in the way the audience hears it*’ quotes Professor Ingo Kock, Dean of the Faculty of Sound at the Potsdam-Babelsberg Film University [The Tagesspiegel, April 17, 2016 / No. 22 731]. The idea is striking, but younger generations will not be pleased with the result. What are the causes of this and what are the solutions? To do justice to the claim of inclusion for all generations of listener, it is necessary to analyse the overall situation and match the solutions to it.

Speech intelligibility of broadcast audio depends on many factors. At the broadcaster, recording, processing, mixing and deficiencies in the transmission chain all play a part. At home, the receiver technology, the acoustic conditions in the listening room and the individual hearing characteristics of the listener all have a bearing. The problem is very complex (Figure 1) and the causes are diverse.



**Figure 1:** Transmitter-receiver model: The causes of poor speech intelligibility are manifold and can be located at different points in the production/transmission chain

Despite various recommendations, guidelines, directives and publications, complaints continue to arise. Audience reactions thereby reflect a wide range of individual auditory impressions.

We will examine several factors that can influence sound quality from a programme's conception to its consumption at home, looking at the various technical parameters set in post-production and broadcasting that affect the transmission quality and recognising that ultimately, speech intelligibility is as individual as consumers' living rooms and their hearing characteristics.

Speech intelligibility should not, however, be considered solely from a technical point of view; is the lack of intelligibility due to modernity? Is it due to technology or to poor speaker/performer training?

It can be because of content and design decisions that are made in the context of a TV feature film production, and which can lead to speech intelligibility issues but that are nevertheless part of an artistically coherent work.

Speech comprehension is a highly complex process with numerous influencing factors. Audio is traditionally recorded *subjectively*, whether by sound colleagues on set or in later production steps in the studio. A current project that aims to develop an *objective* measurement tool for speech intelligibility is also discussed.

## **2. Production and Pre-production**

The issue of *poor speech intelligibility* has ranked number one on the 'negative hit-list' of viewer complaints for quite some time. Speech intelligibility problems are not exclusively a sound engineering issue; they also depend on production, editing, directing decisions and financial constraints.

Over the last decade or so, the production rates per minute paid by broadcasters have fallen across all TV formats, especially in TV series and documentaries. According to [2], in 2012 an average of 1.43 M€ was still available for an episode of the long-running German/Austrian/Swiss crime television series 'Tatort'. In 2016, this average had dropped to around 1.27 M€. The number of shooting days allowed for an episode had also decreased from an average of 28 to 23 days.

The sound budget has also been affected by these cuts. The use of a second boom operator is now almost a thing of the past. Instead, shooting with two and more cameras causes the background noise level for the scenes to increase: Everyone who can be seen in a long shot must also be able to act and be heard. The boom is pushed back, replaced by wireless clip-on microphones – a practical solution on the set, but a mediocre compromise from a sound engineering point of view.

## 2.1 Awareness

In 2014, the ARD and ZDF Television Operations Conference issued the *Speech Intelligibility in Television* recommendations for programmes and technology [1]. It says ‘Sound should be considered as a central part of the programme at all stages of production – from planning to final mixing, because it plays an essential role in storytelling’. Jonathan Pauli presented an interesting thesis on this subject in 2011 [3]. According to Pauli, the design of the sonic world has as much influence on the visual design as vice versa. Pauli's goal is to create a corresponding awareness of the synergies between image and sound, and thus also a basis for ‘improved, holistic, audiovisual film design’.

## 2.2 Pre-audibilisation

Pauli questions the established post-production workflow and proposes the principle of ‘*pre-audibilisation*’ (pre-aud for short), according to which the sound designer produces a preview in the pre-production phase. From script excerpts and ideas resulting from spotting sessions (together with the director and the camera team), an overall audiovisual concept of the film can be created in the form of a ‘storyboard’ wherein all planned elements of the soundtrack (language, sounds, effects, atmosphere, music, ...) are entered along the timeline and are visible to all departments.

Pauli: ‘For the design of the soundtrack, the purpose of the storyboard is to give an overview of the density of the elements and possible dynamic progressions.’ A lower density of auditory elements may well help the audibility of the mix, and thus also help speech intelligibility. The storyboard also points out *dramaturgical condensations* and allows ‘a planned coordination and shaping of the elements of the soundtrack along the filmic dramaturgy’. For the overall production planning, this makes it easier to detect potential problems regarding the choice of location, the ordering of the right equipment, well-qualified personnel and the appropriate time budget, and to prepare accordingly.

## 3. At home with the equipment

According to various studies, in Germany alone, around 12-14 million people between the ages of 15 to 75 suffer from hearing loss that requires remediation [4]. The average age of viewers of public broadcasting is around 60 – German broadcasters 3sat, ARTE and Phoenix serve viewers in their mid-fifties, the private stations have a target group between 45 and 50. Pro Sieben is by far the youngest station, with viewers averaging around 35 years old.

With age, hearing limitations become common: about 25% of 50- to 60-year-olds, 37-50% of 60- to 70-year-olds and over 60% of those over 70 are hard of hearing. The idea of leaving the mixing of movies to seniors over 70 is certainly striking, but

not appropriate to allow for artistic freedom, new forms of expression and audiophile zeitgeist.

### **3.1 Listening environment**

The room and monitoring environment have a major influence on speech intelligibility; different people in different monitoring conditions have very different listening preferences [5]. The listening volume, the playback acoustics with possible interference levels (open windows, children's screams, other background noises), the quality of equipment and the different playback formats used all have an impact.

The demands of seniors and young people regarding an audio mix are rather different, even without hearing loss. Even though most consumers still prefer TV at home, there is increasing consumption – especially by the younger generation – of live streams and media libraries, by means of smartphones, tablets, PCs or laptops. To meet the claim of inclusion for all generations in any listening condition, target group-specific mixes and personalisation of the transmission are quite conceivable.

### **3.2 User meets technology**

Nowadays, most living rooms have flat screen TVs. With them are used home theatre sound systems, which are supposed to compensate for the rather mediocre reproduction properties of the flat screen loudspeakers. But the necessary know-how to best set them up does not always meet their technical possibilities.

Newer systems offer a wide range of setting options that can, under certain circumstances, also make the dialogue less clear. Critical user settings can be found, for example, in the context of 'Room Simulation Features', which add artificial reverb to the signal, or in 'Bass' & 'Bass Boost'/'Treble'/'Equalizing'. These spectral changes can have both positive and negative effects on intelligibility. In combination with spectral colourations already made during recording or post-production (e.g., to achieve a 'crispy-fresh' sound image), compression at many points in the production chain, special transfer functions of TV speakers and individual room sounds, the end of the entire transfer chain cannot be overlooked during production. In addition, 'spatial sound features' can possibly cause phase shifts between the side channels, which lead to an attenuation of the signal part that is common to both sides – usually speech – and thus have a rather negative effect on speech intelligibility.

When using home theatre systems, depending on the interfaces used (HDMI, SPDIF, etc.) and because of the ways that different equipment chains interact: Is the volume changed on the TV set or on the home theatre system or on the receiver? How can inputs and outputs of the device chain be 'calibrated'? Different Dolby reference levels for signal normalisation do not make the behaviour of home theatre receivers particularly transparent to the layman. Many set-top boxes (STBs) and TVs with built-in receivers (IDTVs) still lack the loudness alignment between 'PCM' and 'Dolby Digital'. All contribute to problems in intelligibility, which the layman does not

necessarily grasp. The solutions to these problems were published by the EBU some years ago. [6].

#### 4. Postproduction and transmission

Although production in stereo is still common for television use, multichannel mixes are gaining more ground, culminating in immersive surround sound environments. Whether 'hand-crafted' or supported by up-mixing software, backward compatibility with stereo/mono listening should not be forgotten or possible artifacts should be taken for granted. The 'divergence' parameter is very often used in shows, sports and documentaries. Here, the centre (dialogue) signal is added to the adjacent front channels by means of a potentiometer, with the aim of making sound sources (including dialogue) broader and less direct.

In the multichannel mix, the loudness usually remains constant regardless of the strength of the divergence, but in the downmix there can be loudness deviations of up to 3 LU between the multichannel mix and the ITU-configured downmix. In addition, further disturbing effects occur at the listening position. Due to the overlay of the centre speaker and the phantom sound image via the left and right front speakers, an offset of the (ideal) listening position by only a few centimetres results in shifted phase and frequency responses, which have a significant effect on the sharpness of the sound image and most likely also influences speech intelligibility. More user studies are needed to investigate this.

##### 4.1 The power of Metadata

Metadata is data that is additionally set in the Dolby Digital stream and that has a descriptive as well as a controlling function. Among other parameters, the consumer metadata *Dialnorm* and *Downmixing* parameters play a significant role in the consumer's listening experience. Incorrectly set metadata can have negative effects on sound quality and speech intelligibility for the listener.

The *Dialnorm* metadata determines a level shift at the decoder with the aim of mapping the loudness of the individual programmes at the receiver side. If measured properly, an accordingly set *Dialnorm* parameter should result in reliable listening levels across different programmes and also adequate use of the Dynamic Range Control profiles.

*Downmixing*: In media players (such as set-top boxes, TV or DVD and Blu-ray devices) the analogue stereo output is usually controlled by one of two downmix variants of the Dolby Digital stream: One variant is the *Pro Logic* or *Left total/Right total* (Lt/Rt) downmix, the other is a simple mono-compatible stereo playback, called *Left only/Right only* or Lo/Ro, most appropriate for headphones or dedicated stereo equipment. The difference between the two downmix variants lies in the mixing of the individual channels: The Lt/Rt downmix sums both surround channels and adds the result in-phase to the left stereo channel and out-of-phase to the right. With the



Lo/Ro downmix, the respective surround sides are simply added to the respective stereo channel. The LFE (Low Frequency Effects) channel is not included in the downmix in either case.

## **4.2 Metadata meets technology**

The Lt/Rt format was originally developed to allow an upmix of two channels to multichannel surround sound. Cinema operators had a kind of analogue backup in case the digital multichannel source failed and consumers at home could enjoy surround sound despite the availability of just a two-channel source. Even though Lt/Rt is traded as being stereo compatible, it results in a different stereo image than Lo/Ro. Negative side effects of Lt/Rt can be a rather hollow sound and an unstable, unrealistic positioning of single sound elements. The Lt/Rt format is no longer up-to-date and has been replaced by dedicated surround formats such as Dolby Digital. Nevertheless, it is still frequently used by default.

Which of the two systems that finds its use in home equipment is unpredictable. Theoretically, the broadcaster can set the 'preferred downmix' metadata to one of the two options. However, a study conducted in the Netherlands in 2015 [7] showed that most receiver devices simply ignore the set downmix preference. Some devices allow the user to choose between Lt/Rt and Lo/Ro within the menu settings, but the typical user is rarely able to use this consciously and profitably. More often, however, the downmix scheme turned out to be predetermined by the manufacturer and not customisable within the device settings at all and the use of the Lt/Rt or the Lo/Ro downmix depended on the equipment manufacturer, brand and model. For broadcasters and studios, this means that both downmix variants, Lt/Rt and Lo/Ro, must be checked and monitored for quality during production. Lt/Rt and Lo/Ro can lead to different mixing ratios, on which speech intelligibility primarily depends.

## **4.3 The Downmix of the Upmix**

The format in which television is transmitted, whether stereo, surround or both using simulcast, is up to the broadcaster. If a programme is produced in stereo only, an upmix is required for simulcast. The performance of various upmixers varies significantly and is strongly dependent on the selection of the upmix settings used. It is important to note that the provided multichannel services are not only played back via surround equipment, but also via stereo TV sets. Here again, problems that are dependent on the TV brand and model become apparent: Some of the devices use the Dolby Digital multichannel service by default, even if a true 'hand-crafted' MPEG-1 Layer II stereo signal is available. This in turn leads to a large part of the audience always hearing the stereo downmix reproduction instead of the dedicated stereo mix, usually without realising it.

If end devices do not reliably support assignment to the corresponding Lt/Rt and Lo/Ro formats, the demand for acceptable downmix compatibility from the



manufacturers of upmixers turns out to be problematic. In view of this, the major broadcasters in the Netherlands have adopted the following settings: ‘Preferred downmix = Lo/Ro’ AND ‘90 degrees phase shift = enabled’. If the ‘preferred downmix’ is set to Lo/Ro, the phase shift hardly has any negative effect and loudness differences between surround and downmix are reduced. In case the end device forces the Lt/Rt downmix, which is often the case [7], the ‘90 degrees phase shift’ reduces the typical drawbacks of the Lt/Rt downmix.

For German broadcasters, phase shift is deactivated by default – fair for Lo/Ro, bad for Lt/Rt. More extensive tests have shown that the Lt/Rt downmix from the upmix leads to strange effects in the audio transmission, which also negatively affect speech intelligibility. The Lo/Ro downmix leads to much more stable and reliable results.

In the Netherlands, individual broadcasters discussed whether simultaneous stereo and multichannel broadcasting via simulcast should be discontinued. The decision was made in favour of continuing simulcast. Nonetheless, some distributors only support Dolby Digital sound, a consequence of the general battle for bandwidth (ADSL, satellite). The consequence is that in many cases only upmix versions of stereo productions arrive in living rooms, which in turn are converted into stereo reproduction by the receiving device via downmix.

The broadcasters are thus increasing the pressure on distributors to continue simulcasting or to at least broadcast dedicated stereo audio instead of multichannel. According to Florian Camerer (ORF), a similar decision is also pending in Austria.

#### 4.4 Loudness and side effects

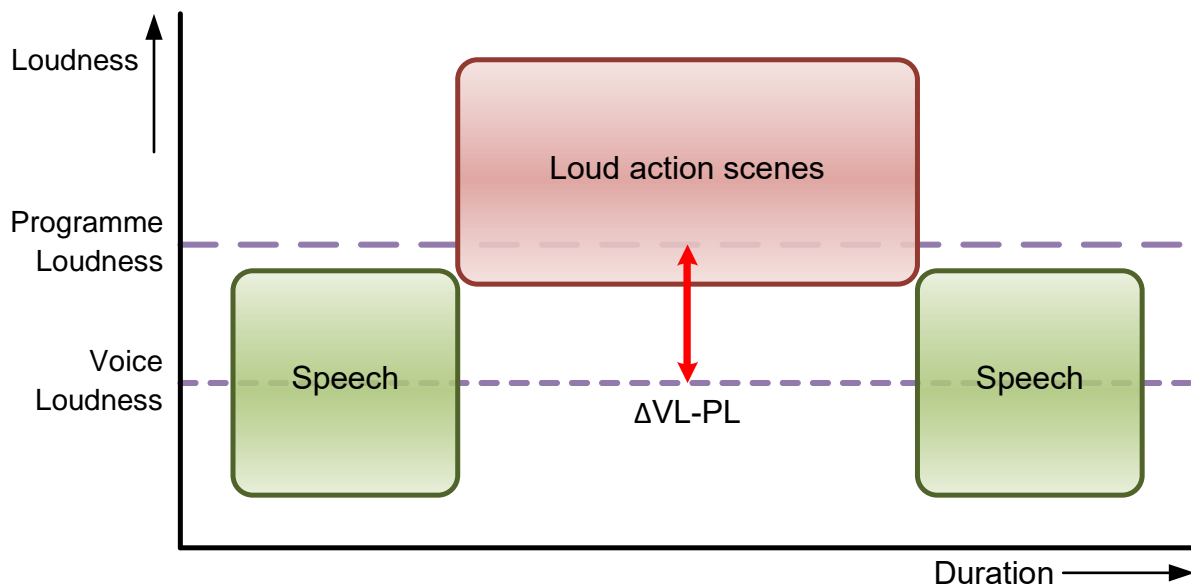
Since mid-2012, public and private TV stations in Germany have normalised their programming according to the EBU R 128 loudness recommendation, with the intent to broadcast with uniform loudness. Despite great successes with regard to the problem of loudness jumps between stations and programmes, weaknesses of the EBU methodology were also identified – especially with particularly dynamic material.

**Table 1:** Comparison of television programmes (samples) with regard to Programme Loudness and Dialogue Loudness (Source: Uwe Krämer)

Programme	Parameter		
	Programme Loudness [LU]	Dialogue Loudness [LU]	Loudness Range [LU]
Tagesschau (daily news)	0	0	4
Fast & Furious (film)	0	-4	20
Gattaca (film)	0	-2 to -15	20
Die Dolmetscherin (film)	0	-4 to -6	16
Mission Impossible (film)	0	-5	15
Tatort (crime TV series)	0	-3 to -5	14

In the latest edition of the *Practical Guidelines for Loudness Normalisation*, EBU Tech 3343 [8], an extra chapter is dedicated to feature films. The concept of Voice Loudness (an equivalent to dialogue level) was introduced and a supplement to normalisation was considered, which is additionally based on an ‘anchor signal’ such as speech, a kind of ‘signal type gating’: loudness is only measured where speech is present. The difference between Voice Loudness (VL) and Programme Loudness (PL) is generally increasingly noticeable in material with a high loudness range. And similarly, in movies and series where the proportion of action scenes (rather loud and effects-laden) is dominant relative to dialogue scenes (rather quiet, corresponding to Voice Loudness).

Both descriptions best fit to the genre ‘feature film’. Conceptually, the Programme Loudness of the complete film is, as agreed, normalised at  $-23.0$  LUFS/0 LU. However, if one measures the dialogue during the film, they can differ considerably (cf. Table 1).



**Figure 2:** Loud action scenes dominate the calculation of Programme Loudness. Proportionally, a difference between Programme Loudness and Voice Loudness exists

This has consequences for the viewers. If, for example, they are watching the daytime news and then, in the subsequent feature film, ‘Mission Impossible’, the dialogue of the latter is too quiet and difficult to understand at the currently set listening volume. Re-adjusting the volume is necessary despite the loudness revolution. However, if the speech volume is set sufficiently loud, the effects of the movie may often be perceived as too noisy – for the viewer, this problem falls into the category ‘speech intelligibility’.

## 4.5 Special normalisation for film material

Another study in the Netherlands [9] led to an automated process specifically aimed at normalising film and series material for television. To establish this, a maximum difference of 4.5 LU between Voice Loudness (VL) and Programme Loudness (PL) was defined, which is adjusted accordingly with the help of audio software. Movies and series are subsequently normalised with respect to their VL. Mixes where the difference between VL and PL is less than 4.5 LU are left unchanged by the software in their dynamics, but the Voice Level is nevertheless increased to -23 LUFS. This adjusts the speech volume to the usual dialogue loudness of the rest of the transmission to avoid loudness jumps during, for example, commercial breaks.

In cinema mixes, where differences between VL and PL of up to 14 LU are quite common, these are automatically reduced and adjusted in their dynamics, as far as possible without limiting the quality of experience and the original intention of the dramaturgy. As a result, the Programme Loudness of feature films and similar television series on TV is allowed to exceed the usual -23 LUFS by 4.5 LU – a kind of exceptional permission.

This software has been used successfully in the Netherlands since 2014, so far without any viewer complaints. The methodology was recognised as ‘highly commended’ in the 2015 IBC Innovation Awards ballot [10] and is now being considered as an official extension to the EBU R 128 recommendation.

## 4.6 The Pre-Emphasis Bit

By boosting high frequencies (pre-emphasis), the signal-to-noise ratio of a transmitted signal can be improved. These signal adaptations are normally reversed during reception for faithful reproduction. However, if the status of the pre-emphasis bit is incorrect, the listener experiences a significant degradation in quality. For example, if the high frequencies in the audio signal have not been boosted, but the pre-emphasis bit has been set in such a way that the terminal equipment reverses the suggested pre-emphasis, there will be a significant attenuation of the high frequencies, in other words, a musty sound with poor speech intelligibility. Regular sweep tests to check the transmission chain should help detect such mis adjustments.

Similarly poor results in the sound image can of course also come from a malfunction of the individual decoder, such as a specific model of TV or set-top box.

## 5. Slight blurring of the tone...

Image design has changed since the early days of film. Black and white became colour, silent film became talkies, everything changes with time and experience, especially when creativity and technology meet – shot sizes and visual aesthetics, film stock and contrast levels, resolution of a scene and editing frequencies. As

technique and performance have changed, so has the viewer's perception and understanding of narrative structures. New montage concepts have established themselves, others have yet to make the breakthrough – viewing habits are subject to the zeitgeist and thus to change.

Sound and its editing have also changed. From mute to mono, to stereo, to discrete multichannel and immersive audio. Listening habits are also subject to the zeitgeist. Is poor speech intelligibility perhaps also a stylistic tool? Comparable to a slight blur in a picture? Or reality? Even in real life, the desire for understanding dominates (the basic laws of perception psychology are not subject to the zeitgeist) but is not always fulfilled. In the following, dramaturgical, content-related and creative decisions will be illuminated, which can ultimately lead to speech intelligibility criticism.

## 5.1 The problems start before the film is shot

Before sound designers even get a chance to work, the basic production parameters are defined by the programme managers (editors):

*What content (script) is being filmed?* Depending on the content, problems of comprehension may already arise, for example, in films with a high proportion of informative dialogue in a noisy environment or very complicated, convoluted stories.

*Who is responsible for the design of the screenplay?* The creative realisation depends to a decisive degree on their tastes and the artistic freedom they are given. If the director only marginally cares about speech intelligibility and does not accept corresponding suggestions from the sound experts, the die is practically already cast.

*Which role will be cast with which actor?* If, for example, the speaking style of an actor is known to be mumbled, soft and fast, this will rarely be revised in the further course of the production.

The people responsible for sound post-production usually have only two options.

One is to 'escape to the front' – in other words, to continue working creatively in the sense of the concept and to consistently follow the paths taken on the set. This leads to a coherent overall product, which can, however, come under criticism in terms of speech intelligibility.

The other is to go into 'confrontation' and refuse to realise the film creatively in this way, try to arrange for a lot of dubbing (if necessary, by other actors) and/or advocate the partial re-cutting or exchange of scenes. The result would be longer editing time, higher costs for the client, and possibly a cinematic work that no longer appears to be of one piece, since sound-language and image-narrative no longer fit together. Possible approaches for improvement would be:

*Preliminary clarification:* Mixing engineers and sound designers are involved in the planning and realisation process right from the start, to develop a procedure together

with everyone that leads to good speech intelligibility (cf. § 2.1 Pre-audibilisation). This procedure is conceivable in the television production process but is unlikely.

*Compulsion:* A legally binding obligation to produce fundamentally accessible feature films or at least to produce a second completely accessible version, could be enacted. This is also possible in principle but would likely result in enormous costs for the client, since this would not be feasible through mere level changes, and it would require extensive re-editing. The work would largely have to be changed.

## **5.2 Speech intelligibility is not the same as film intelligibility**

In the past, television feature films were usually clearly directed, acted, shot, edited and, of course, spoken. Poor speech intelligibility was rarely an issue. Many of today's productions have elements borrowed from cinema, such as high dynamics in image and sound, fast cuts, off-screen action, 'authenticity' and complicated, thematically difficult stories that can cause problems in intelligibility even for viewers who are not hard of hearing.

Programme makers are caught on the horns of the dilemma of including as many viewers as possible, and of providing artistic and cultural stimuli. Feature films should stand out from the mass of productions and should attract attention and generate social relevance. It is inevitable that in certain cases this may result in poor speech intelligibility because new creative paths are being taken. Whether or not such 'difficult-to-understand' productions should be shown on television is clearly in the hands of those responsible for programming – not those responsible for sound.

On the other hand, the task of sound supervisors to produce a sound that matches the picture has hardly changed over the years: It is not a matter of creating a sonically neutral image of the situation on set, but rather a creative-artistic process in which speech is only one design element among many. Human communication takes place not only in words, but also on other levels:

- Verbal level: The content of the spoken word.
- Paraverbal level: The manner of speaking, such as shouting, whispering, crying, etc.
- Nonverbal level: Body language, such as gestures, facial expressions, posture, gait, head movement, etc.



**Figure 3:** German crime scene: Tatort, ‘Der irre Iwan (The lunatic Ivan)’ [14] – In the background, a fair, one character in a rabbit costume, an enraged (Austrian) female character. Non-verbal and para-verbal intelligibility are excellent, verbal intelligibility is unsatisfactory

Artistically rich acting takes place equally on all three levels and is reproduced both in film and television. There are perceptual-psychological findings that the nonverbal accounts for well over half of a person's total expression. This is another reason why only part of the attention is paid to verbal intelligibility in the production process. The paraverbal part is transmitted more easily in sound, as it works even at lower levels or amongst other interference. The non-verbal part is transmitted in the image.

A measuring device that only measures the purely verbal comprehensibility cannot make a statement about the comprehensibility of the complete cinematic expression, and thus does not meet higher creative demands.

### 5.3 Not all television is the same

The *Technical Guidelines for the Production of Television Productions for ARD, ZDF and ORF* of April 2015 [11] state on page 29: ‘The sound recordings must correspond meaningfully with the picture in terms of design. They must not contain any unintentional changes to the acoustic atmosphere and must have a balanced mix ratio throughout. For a version suitable for television, the mixing ratio must always be selected in favour of speech intelligibility.’

The fact that design standards are clearly specified in technical guidelines is evidence of the not necessarily promising attempt to establish uniformity in television design. However, the spectrum of content ranges from news, documentaries, sports, entertainment, art, music, political discussion to feature films. One distinguishing criterion here can be the degree of artistic freedom, which, for example, is

considerably less in the German daily news programme 'Tagesschau' than in the 'Tatort' drama series.

The more artistic freedom there is, the less is the standardisation that is possible. That speech intelligibility should be assumed to be the ultimate goal for all productions can be considered debatable: In non-fiction programmes, standardisation could certainly make sense. However, it is the case that in strongly artistically oriented films, controversial opinions can arise among the viewers or even be provoked to initiate a discourse, and this is anchored in the nature of art.

## 6. Measure Speech Intelligibility

The speech intelligibility of spoken contributions is traditionally recorded *subjectively*, whether by the sound colleagues on set or in later production steps in the studio. Within a production process, recordings are usually listened to many times in an optimal listening environment – so it is easy to misjudge mix ratios and their speech intelligibility. In post-production, in addition to the viewing, editing and processing of the previously recorded material, various steps of dubbing and final mixing take place. For this final mix, several audio tracks and audio elements such as music, original sound, atmosphere, noises, voice-overs, etc. flow together, which can also lead to changes in speech intelligibility.

### 6.1 Target group-specific measurement

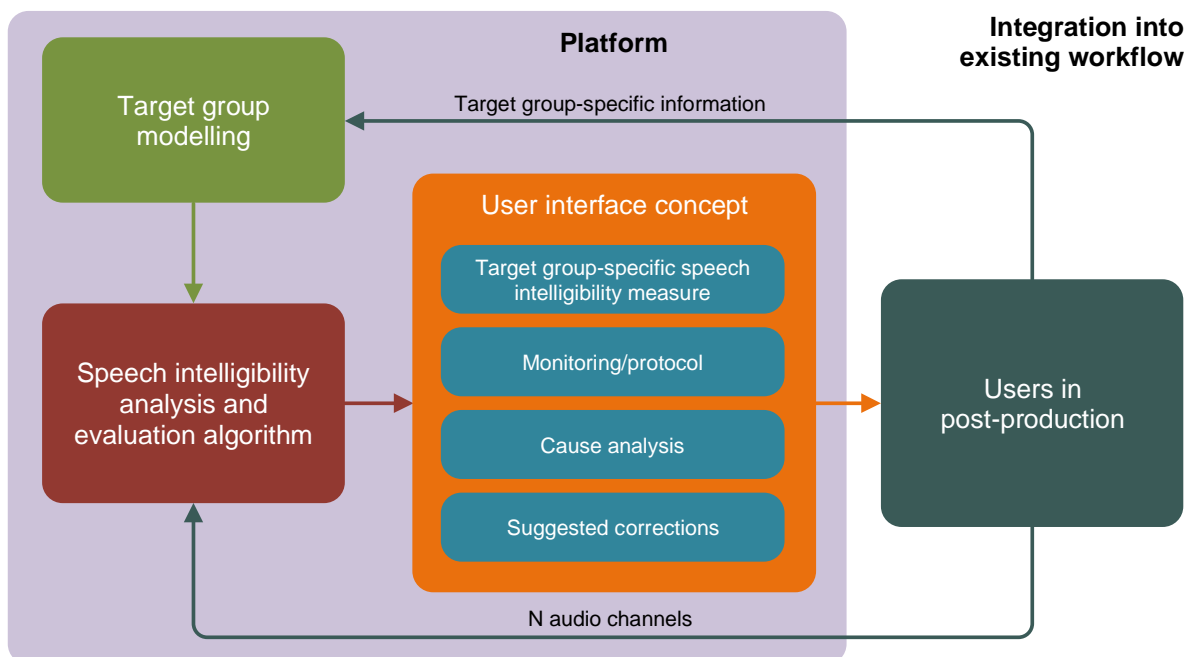
It is impossible for sound engineers, editors and broadcast managers to adequately take account of the specific requirements of different user groups (in terms of good speech intelligibility) for production and, above all, for mixing, without an appropriate tool. In view of the large number of parameters to be observed and monitored by those responsible for the sound, a display of speech intelligibility values that can be quickly recorded and clearly interpreted and possibly combined with recommendations for action, would provide relief. Especially at the workplaces of journalists, editors and cutters, where sound is not explicitly the focus of the editor, a user interface that is as simple as possible and suitable for laypersons, would be of interest.

Against this background, the **SI4B** – Speech Intelligibility for Broadcast – project came into being in 2016. The goal of the SI4B project [12], which is funded by the BMBF (the Federal Ministry of Education and Research in Germany), is to develop latency-free algorithms for continuous and objective monitoring of speech intelligibility and its representation. Spectral, temporal and energetic properties of the speech and background signal are evaluated to model and display the intelligibility of speech, especially within the sound mix.

With the help of this display, speech intelligibility can be monitored. The speech intelligibility monitoring also offers the possibility to check sound mixes for target group-specific speech intelligibility or to produce mixes for special target groups.



Within the project, the two target groups '*normal hearing*' and '*hard of hearing*' are addressed as examples.



**Figure 4:** Schematic representation of the speech intelligibility measurement system

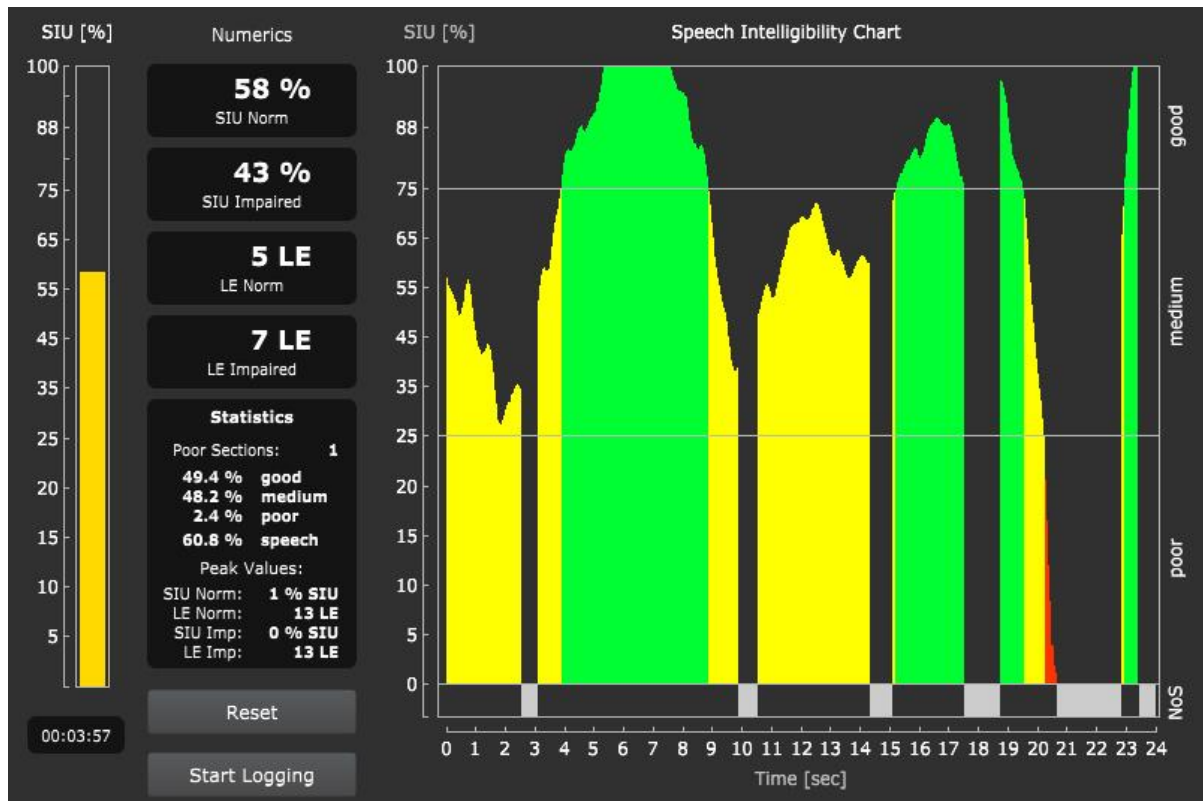
Figure 4 shows a schematic representation of the system. The audio tracks to be evaluated (speech, music, original sound, atmosphere, effects, etc.) are mixed by the engineer in post-production as usual and fed into the analysis module as audio signals. The analysis module evaluates the speech intelligibility of the mix and provides analysis results as transfer parameters (red arrow) for the presentation concept. If desired, the user can specify information about the target group (e.g., age, typical hearing ability, listening conditions) to perform the modelling of speech intelligibility in a target group-specific way.

## 6.2 Control and visualisation module

The number of distribution channels to be monitored is constantly increasing. At the same time, however, there is a shift in production work towards non-technical or non-technically trained operators, such as editors. All of them must be presented with a secure representation that is unambiguous in its interpretation, so that a product that is perfect in terms of speech intelligibility can be created or distributed. The user interface must represent the following facts in this context:

*Target group-specific speech intelligibility measure:* The displays in the user interface are switchable to the respective target group, if required. Parallel observation of target groups is also possible. An appropriate measure for the evaluation is presented in Speech Intelligibility Units (SIU). A value in the range 75-100 SIU then covers sufficient to good speech intelligibility.

*Monitoring/ Protocol:* The assessment and display for degrees of speech intelligibility is transformed into an intuitive and meaningful real-time display form and evaluated (cf. Figure 5).



**Figure 5:** Measurement along the time axis [15]

In addition, the module allows logging of the speech intelligibility measure over entire posts. A dynamic (autozoom) highlighting of critical sections facilitates the identification of problems for the sound engineer. The weighting of the degree of interference, duration of interference, and frequency in the speech intelligibility assessment offers a representation that dynamically adapts to these characteristics.

*Root cause analysis:* The causes for the deterioration of speech intelligibility can be very complex. The objective is to show the user/sound engineer the probable causes of poor speech intelligibility. To achieve this goal, the algorithms are used to evaluate which input parameters in the respective measurement leads to poor results in the qualifier. Since the qualifier will likely operate in multiple stages, this investigation must be performed at each stage. A simulation may also be necessary to evaluate which parameters need to be suitably changed to improve the result of the qualifier.

The development of the speech intelligibility measurement tool is now advanced [12] and is already being used in individual products. Also, automated processing of the mix towards an additional, more easily understandable mix is close to entering the market. More about the novelties of this development can be found in [13].

## 7. Conclusions and appeals to stakeholders

The variety of causes for poor speech intelligibility at the receiver is complex – this article could certainly only deal with a part of the effective influences. It is already apparent that there can be no ‘universal panacea’. The discussions at the VDT convention have shown interesting technical conditions and interrelationships that are difficult to grasp, even for professionals.

One appeal, for example, goes to the equipment manufacturers to ensure that the devices at the consumer end, process the downmix metadata in a consistent and meaningful way.

An appeal goes out to broadcasters and network operators to push for regular monitoring of broadcast paths and end results.

An appeal goes out to the sound and programme managers to ensure qualitative recordings and mixes and to make these possible.

A measurement procedure to objectify the traditionally subjective parameter of ‘speech intelligibility’ makes it easier for those responsible for sound to assess the production, especially for diverse target groups, and strengthen the dialogue with directors and editors. However, it does not seem to be sufficient to check the quality in production only, since there can still be quality losses in distribution. In this respect, an appropriate quality measurement and improvement device at the recipient's location would certainly be desirable.

Last but not least, an appeal goes out to all viewers to ensure adequate room acoustics and sound reinforcement in their living rooms. If each party does what it can, modern mixes or even ‘art at the crime scene’ should continue to be allowed. More so if future technical means could allow the production and dispatch of target group-specific or even individualisable mixes.


## References



- [1] ‘Speech Intelligibility in Television - Recommendations for Programmes and Technology’ (*‘Sprachverständlichkeit im Fernsehen – Empfehlungen für Programm und Technik’*), ARD/ZDF, Auftraggeber FSBL-K, 2014.
- [2] Castendyk, O., Goldhammer, K.; ‘Producer Study 2012. Data on the Film and Television Industry in Germany 2011/2012.’ (*‘Produzentenstudie 2012. Daten zur Film- und Fernsehwirtschaft in Deutschland 2011/2012’*), Forschungs- und Kompetenzzentrum Audiovisuelle Produktion an der Hamburg Media School, Hamburg (Hrsg.) Berlin 2012.
- [3] Pauli, J. (2011), ‘Sound Design in Pre-production’ (*‘Sound-Design in der Vorproduktion’*), Hochschule der Medien Stuttgart, Masterarbeit im Studiengang Elektronische Medien.
- [4] Sohn, W. (2001), ‘Hearing loss in Germany, representative hearing screening survey of 2000 subjects in 11 general practices’ (*‘Schwerhörigkeit in Deutschland, Repräsentative’*

*Hörscreening-Untersuchung bei 2000 Probanden in 11 Allgemeinpraxen*) in: Z. Allg. Med. 2001; 77; 143-147, Hippokrates-Verlag, Stuttgart.

- [5] Rennies, J., Oetting, D., Baumgartner, H., Appell, J. (2016). 'User-interface concepts for sound personalisation in Headphones', 2016 AES International Conference on Headphones.
- [6] EBU Tech 3344, 'Guidelines for distribution and reproduction in accordance with EBU R 128', Geneva, 2016.
- [7] Research in the Netherlands performed by Richard van Everdingen at SBS Broadcasting, RTL Netherlands and NPO (Nederlandse Publieke Omroep).
- [8] EBU Tech 3343, 'Guidelines for production of programmes in accordance with EBU R 128', Geneva, 2016.
- [9] Research in the Netherlands performed by Richard van Everdingen at SBS Broadcasting / Talpa Network.
- [10] IBC Innovation Award 2015 finalist for Content Management, <https://www.tvbeurope.com/ibc/broadest-range-yet-ibc-innovation-awards>.
- [11] TPRFHDTV 2014, 'Technical guidelines - for the production of HDTV television productions for ARD, ZDF and ORF' (*Technische Richtlinien – HDTV zur Herstellung von Fernsehproduktionen für ARD, ZDF und ORF*), publisher: Institut für Rundfunktechnik, as of April 2015.
- [12] Cooperation project 'Objective analysis, visualisation and correction of speech intelligibility in broadcast applications for normal and hard of hearing people' (*Objektive Analyse, Visualisierung und Korrektur von Sprachverständlichkeit in Broadcastanwendungen für Normal- und Schwerhörende*), funded by: Federal Ministry for Economic Affairs and Energy based on a resolution of the German Bundestag.
- [13] Hannah Baumgartner; 'Crime scene speech intelligibility: neural networks in audio processing and evaluation' (*Tatort Sprachverständlichkeit: Neuronale Netze in der Audioverarbeitung und Bewertung*), publication date Oct 7, 2020 publication description Verband Deutscher Tonmeister - VDT live, <https://tonmeister.org/de/termine/vdt-live/beitraege/>.
- [14] Huber, Richard (2014), 'Tatort', episode 929, 'Der irre Iwan', Weimar Filmproduktion/Wiedemann & Berg Television, screen capture taken from ARD Mediathek [TC 1:20:49], originally transmitted by Mitteldeutsche Rundfunk (MDR) on January 1<sup>st</sup>, 2015. © W&B Television GmbH Germany.
- [15] Speech intelligibility meter, prototype implemented by RTW. © RTW GmbH & Co KG Germany.

## Author(s) biographies

	<p>Hannah Baumgartner has a master's degree in Hearing Technology and Audiology from the Carl-von-Ossietzky University in Oldenburg. She is trained as a Media Designer for Image and Sound (Mediengestalterin Bild und Ton). Since 2013 she has worked as a research assistant at the Oldenburg institute branch HSA (Hearing, Speech and Audio technology) of the Fraunhofer Institute for Digital Media Technology. There she dealt with the possibility of objectively measuring and displaying the speech intelligibility of audio mixes. She has worked as a sound engineer and camera assistant for over 10 years and wrote as a freelance journalist for the German magazine 'Film and TV Camera' (Film und TV Kamera). Since 2017, she has also been an active member of the board of the Association of German Sound Engineers VDT (Verband Deutscher Tonmeister).</p>
	<p>Richard van Everdingen has worked as an Information and Communications Technologies computer engineer and as a head-end systems specialist for Dutch cable operator Casema. He has patented a measurement system for FM modulated broadcasting, developed a loudness-based levelling system for rebroadcast use and introduced the concept of a levelling system for DVB distribution operating in the MPEG domain.</p> <p>After founding his own company, Delta Sigma Consultancy, he researched for major Dutch broadcasters SBS Broadcasting/Talpa Network, RTL and NPO. For SBS he successfully developed improved loudness normalisation based on a controlled combination of two measurements: full-mix integrated and speech loudness, described in § 4.5 of this article. This breakthrough was nominated for the 2015 IBC Innovation Award.</p> <p>For the Dutch TV market, he led a project group for the application of Event Triggering in TV broadcasting, creating comprehensive specifications for that purpose. Working for the Society of Cable Telecommunication Engineers, he co-authored SCTE-35 and SCTE-104, the worldwide standards for DTV Programme Insertion of Cueing Messages.</p> <p>Richard has been an active member of the EBU PLOUD group since its launch in 2008, where he also led the distribution subgroup. He has given presentations for conventions, societies and companies at various places in Europe, is co-author of the IBC paper 'Toward a recommendation for a European standard for peak and LKFS loudness levels', is the author of the EBU Technical Review article 'Loudness; don't forget the distribution chain' and of EBU Tech 3344, the EBU R128-related 'Practical Guidelines for Loudness in Distribution and Reproduction'.</p>
	<p>Dipl.-Ing. (FH) Bernd Schreiner was born in 1975 in Germany. He graduated in audiovisual engineering in Düsseldorf and started working as a freelance re-recording mixer and sound designer in the year 2000, working mainly for ARRI Film &amp; TV Production Services and Bavarian Television (BR) in Munich.</p> <p>His credits include cinematic films, advertising and ADR (Automated Dialogue Replacement) productions, but he specialised in mixing features and documentary films for television. In 2021 he also acquired a B.Sc. in Psychology, digging deeper into an understanding of audiovisual perception and the challenges of creative collaboration. Bernd is a member of the Association of German Sound Engineers VDT (Verband Deutscher Tonmeister) and the Professional Association Film Sound BVFT (Berufsvereinigung Filmtön).</p>

	<p>Michael 'Mike' Kahsnitz was born in 1958 in Germany. After school and studies at SSL, 3M Minicom Division, he worked as technical director and sound engineer at Dierks Studios in Germany for more than 20 years.</p> <p>He co-founded his own PA company in 1976 and subsequently the remote recording company Eurosound GmbH, performing hundreds of live recordings with many well-known artists. In 1989 he became a specialist for audio measurement at Audio Precision with RTW in Cologne.</p> <p>He is still working for RTW as a senior director of product management. He worked as a teacher at the school for broadcasting technologies in Nürnberg. He is a member of the EBU PLOUD group and the Audio Engineering Society and has authored various publications and convention speeches.</p>
	<p>Uwe Krämer has a Diploma from the University of Düsseldorf in audio and video technology and a Diploma from ILS Hamburg in Management techniques and corporate governance. He worked for several years as a research assistant at the Institut für Rundfunktechnik and since 1990 as a lecturer/trainer at the ARD ZDF medienakademie (formerly the Schule für Rundfunktechnik,) in Nürnberg, where he is also involved in developing and producing e-learning modules in the producer team.</p> <p>In addition to the basics of recording studio technology (sound processing and sound design), his expertise encompasses multi-channel sound, loudness, speech intelligibility and audio quality control. His very wide-ranging and very practically focussed teaching activities and professional experience has resulted in him authoring/co-authoring many publications and presentations. He participates in numerous broadcasting committees and ad hoc groups, and he is currently the vice-chairman of the Surround Sound Forum.</p>

Published by the European Broadcasting Union, Geneva, Switzerland

ISSN: 1609-1469

Editor-in-Chief: Patrick Wauthier

E-mail: wauthier@ebu.ch

Responsibility for views expressed in this article rests solely with the author(s).