OPEN ACCESS Freely available online
J. Paulus, M. Torcoli, C. Uhle, J. Herre, S. Disch and H. Fuchs, "Source Separation for Enabling Dialogue Enhancement in Object-Based Broadcast with MPEG-H"
*J. Audio Eng. Soc.*, vol. 67, no. 7/8, pp. 510–521, (2019 July/August.).
DOI: https://doi.org/10.17743/jaes.2019.0032



# Source Separation for Enabling Dialogue Enhancement in Object-Based Broadcast with MPEG-H

**JOUNI PAULUS**[1,2], **MATTEO TORCOLI**,[1] *AES Member*, **CHRISTIAN UHLE**,[1,2] *AES Member*,

**JÜRGEN HERRE**,[1,2] *AES Fellow*, **SASCHA DISCH**,[1,2] *AES Associate Member*, **AND HARALD FUCHS**[1]

[1] *Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany*
[2] *International Audio Laboratories Erlangen, Erlangen, Germany, a joint institution of Universität Erlangen-Nürnberg and Fraunhofer IIS*

Dialogue Enhancement (DE) is one of the most promising applications of user interactivity enabled by object-based audio broadcasting. DE allows personalization of the relative level of dialogue for intelligibility or aesthetic reasons. This paper discusses the implementation of DE in object-based audio transport with MPEG-H, with a special focus on source separation methods enabling DE also for legacy content without original objects available. The user-benefit of DE is assessed using the Adjustment/Satisfaction Test methodology. The test results demonstrate the need for an individually adjustable dialogue level because of highly-varying personal preferences. The test also investigates the subjective quality penalty from using source separation for obtaining the objects. The results show that even an imperfect separation result can successfully enable DE leading to increased end-user satisfaction.

## 0 INTRODUCTION

Low intelligibility of narration or dialogue due to too high background level is one of the most common complaints in broadcasting [1]. The underlying reason for low intelligibility may be, e.g., hearing impairment [2], challenging listening environment [3], non-ideal reproduction setup [4], listener's language skill level in the dialogue language [5, 6], or unfamiliar accent or dialect [7, 8]. Even when the intelligibility is not compromised, the personal preference of the listener may differ from the broadcast mix [9, 10].

The problem of low intelligibility can be addressed by providing a second "Clean Audio" track with less background sounds [11]. Producing additional audio mixes requires more resources and it can be prohibitive in some cases [1]. The idea of Dialogue Enhancement (DE) is to provide the end-user with the possibility to adjust the relative level of dialogue to their own preferences and needs without the broadcaster providing multiple mixes [9, 12]. A basic DE functionality can be provided using parametric audio object coding, e.g., [13], but the full potential becomes available with object-based audio supported by the recent broadcast standards, e.g., MPEG-H Audio [14–17].

A challenge for the deployment and adoption of object-based audio transport is that object-based audio production is gradually starting and much of the legacy content exists only as mixes. The end-user may find it confusing or frustrating when the personalization functionality is available only for a few new programs and not for the classics and personal favorites. DE can be implemented having the dialogue as an object separate from the channel bed containing the background. In this paper we *propose a system using a combination of source separation methods for extracting the dialogue content from legacy broadcast audio mix for enabling object-based audio broadcast* with MPEG-H. Even though the source separation result is not perfect, we *show through subjective evaluations that the result is still improving the end-user satisfaction* in the DE application.

The paper is organized as follows: Sec. 1.1 provides a brief overview of MPEG-H. Sec. 1.2 gives an overview of speech separation algorithms that can be used for obtaining the objects. Sec. 2 details the proposed method, and Sec. 3 describes the subjective and objective evaluations. Sec. 4 gives the conclusions of the paper.

## 1 BACKGROUND

### 1.1 Object-Based Audio Using MPEG-H

Based on MPEG Unified Speech and Audio Coding, the MPEG-H Audio standard offers many extensions for use in the context of immersive 3D audio, such as coding and rendering of multi-channel and object signals, transmission of object metadata, the compressed transmission of (speaker layout agnostic) object positions and trajectories, and it allows for personalization and user interactivity on the decoder side that is enabled and controlled by object metadata. The MPEG-H Audio standard was published in 2015 [15], amended in 2017 with the so-called Phase 2 developments and the definition of MPEG-H Audio Low Complexity (LC) Profile [16], and a Second Edition is being issued [17].

The underlying main ideas of the new codec are to provide suitable means for an *immersive experience*, for *universal delivery*, and for *personal interactivity*. The immersive sound experience is provided by supporting 3D loudspeaker setups, adding the height dimension to surround sound, and a binaural renderer provides 3D sound experience on headphones. The universal delivery means that the audio data can be delivered in one universal format and might be automatically rendered in the best possible reproduction mode on the receiving device. Given these two principles, the listening scenarios can be very different, ranging from a reproduction via smartphone and loudspeaker in a noisy city environment up to a high-end speaker setup in a quiet home cinema scenario. By enabling user interactivity the consumers can benefit from adjusting mixing parameters according to the circumstances of their listening situation and to their liking.

In the transport the audio channels might typically be used for a channel bed, while the audio objects can be utilized to enhance the channel bed through addition of user-interactive elements for individual mixing as well as for playback situation based rendering using, e.g., object spatial trajectories conveyed as metadata. The objects can also be controlled individually in terms of their dynamic range, ensuring audibility in all dynamic range compression modes, and they can also be made selectable as alternatives, e.g., different languages, commentary, or audio accessibility aids.

A most demanded use-case is the individual adjustment of the dialogue level over the background music or sound effects. The broadcaster may offer different recommended adjustment presets through object metadata, e.g., in sports scenarios a preset could be “Dialogue+” with a more prominent commentary and attenuated stadium atmosphere, while another preset could be “Stadium” without any commentary. In addition to the presets, the users can fine-tune the relative dialogue level if this is enabled by metadata. A similar setup is also useful for other content types, e.g., TV shows, drama, or documentary, if the dialogue is available as a separate signal.

These object-based use-cases as well as the combination of object-based audio with immersive sound have recently been tested in field trials at the Eurovision Song Contest in May 2018 in Lisbon [18] and in a sports event in Paris in June 2018 [19]. Furthermore, MPEG-H Audio has been adopted in a broad range of broadcast and streaming application standards, such as ATSC 3.0 [20], DVB-MPEG/UHD [21], and DVB-DASH [22], referencing the MPEG-H Audio LC Profile. Moreover, MPEG-H Audio is already on air in a regular service of the terrestrial UHD system in South Korea since May 2017. This system is specified by TTA [23] based on ATSC 3.0.

### 1.2 Speech Separation

Object-based production is gradually starting. More content will become available, while much legacy content will still be re-broadcast. In order to use the object-based transport and enable DE, methods of audio source separation can be applied to split the legacy mixture into estimates of the dialogue and background, as illustrated in Fig. 1. Various methods for separating a target signal from a mixture of signals have been developed in the past. These methods can be categorized into model-based and data-driven approaches.

#### *1.2.1 Model-Based Approaches*

The model-based methods rely on modeling assumptions about the signal or the mixing process. A signal model describes characteristics of the input signals, while a mixing model describes how the input signals are combined to the mixture signal.

Many classical speech enhancement methods belong to the model-based category. The estimation of the noise spectrum using minimum statistics tracking of local minima of the signal energy in each sub-band has been proposed in [24]. A non-linear update rule for the noise estimate and faster updating has been proposed in [25]. Time-recursive averaging algorithms estimate the noise spectrum when the estimated signal-to-noise ratio (SNR) at a particular frequency band is low. The estimation computes recursively the weighted average of the previous noise estimate and the present spectrum. The weights are determined as a function of the speech presence probability or as a function of the estimated SNR in the particular frequency band [26, 27]. Histogram-based methods rely on the assumption that the distribution of the sub-band energy is bi-modal: the low-energy mode for segments without speech or with low-energy segments of speech and high-energy mode for voiced speech and noise [28]. For a comprehensive review of classic speech enhancement methods the reader is referred to [29].

Several methods for enhancing the dialogue in a stereo recording make the assumption that the dialogue is panned to the center, e.g., [30–32]. Other methods with a similar rationale are Azimuth Discrimination and Resynthesis (ADRess) [33] and Degenerate Unmixing Estimation Technique (DUET) [34], where the separation is achieved by binary masking after clustering the time-frequency bins into sets with similar inter-channel time and level differences. Since the target and interfering signals often have similar spatial cues, a separation based on inter-channel cues may leave residual interference in the output signal.

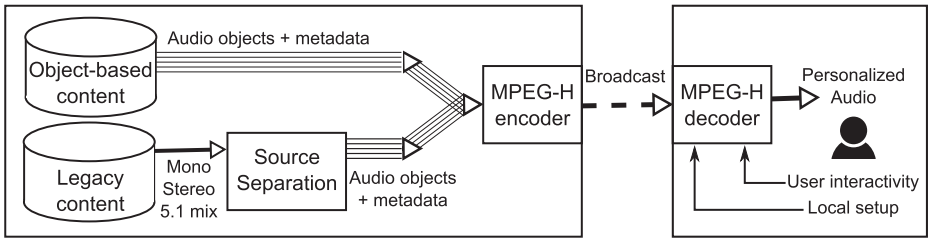


Fig. 1. Full system overview. In the case of legacy content (i.e., only available as mono, stereo, or 5.1 mix) source separation is applied in order to estimate audio objects (e.g., dialogue and background) and enable user interactivity via MPEG-H (e.g., DE).

### 1.2.2 Data-Driven Approaches

In data-driven approaches a representation of the target signal or a set of parameters for retrieving the target signals from the input mixture is estimated. The estimation is based on a model constructed from a set of training data and derived by optimizing a criterion, e.g., by minimizing the mean squared error between the true and the estimated target, given the training data.

Non-negative matrix factorization (NMF) [35] factorizes a non-negative matrix, e.g., a magnitude spectrogram, in a product of two low-rank non-negative matrices, e.g., spectral basis functions and the corresponding temporal activations. Applying constraints on the bases, e.g., temporal smoothness or sparseness, the factorization can be guided to different solutions. NMF-based source separation approaches work mainly on a single-channel input. Semi-supervised NMF uses a fixed dictionary of spectral basis functions for speech, and for these only the temporal activations are subject to optimization. A number of bases are left free for modeling the interferer, and for these both the spectral and temporal bases are optimized. Various ways of constructing the dictionary and applying constraints during the factorization have been discussed in the literature, e.g., in [36] the dictionaries are constructed such that their discriminative power is maximized, while in [37] the spectral basis dictionaries from multiple talkers learned independently are concatenated, and a block-sparsity constraint preferring to use the bases of only a few prototype talkers when modeling an unknown talker is applied. The latter is then improved in [38] by constructing the dictionary to have a higher modeling capability and by using a more strict sparsity constraint.

Another and widely used example of data-driven approaches are artificial neural networks (ANN) that have been trained to output an estimate of a speech signal given a mixture. During the training the adjustable parameters of the network are determined such that a performance criterion computed for a set of training data is optimized. Early approaches with ANNs made a heavy use of elaborated feature extraction [39–41]. The first publication on DE using supervised learning with neural networks [41] processed single-channel input signals and used ANNs with only one hidden layer.

For the processing of signals having more than one channel, spatial information has been employed in [42] for estimating a binary mask with a deep neural network (DNN) having inter-channel features as inputs. A denoising autoencoder using multichannel features has been investigated in [43]. In [44] a DNN with log-power spectra from one channel and channel-level difference has been used for predicting the target signal spectrum, while the log-power spectra of both input channels were used as inputs to a recurrent network in [45]. Binaural processing with single-channel features from a delay-and-sum beamformer output have been used to predict the separation mask in [46]. For a more comprehensive overview on deep learning for speech separation the reader is referred to [47].

## 2 METHOD

The broadcast audio is highly diverse in input signal characteristics with respect to channel format, background type, level, and spatial cues [1], and this is challenging for the source separation. Various source separation methods rely on specific signal characteristics and cannot effectively process all broadcast content alone. We address this challenge by applying multiple source separation methods based on complementary cues in parallel and by combining the results with a late fusion. Particular attention is paid to simple and robust processing of the various channel formats used in legacy content. All input signals are processed as stereo signals: single-channel inputs are converted by duplicating the channel, and 5.1 inputs are processed by applying the speech separation on the center channel and attenuating the other channels (assuming that the center channel carries main portion of the dialogue). Consequently, only the single-channel methods described in Secs. 2.3–2.5 are effective when processing either mono or 5.1 input signals.

The processing takes place in short-time Fourier-transform domain using 21.3 ms frames with 50% overlap corresponding to 1024 and 512 samples at 48 kHz sampling rate, a sine window, and discrete Fourier transform length of double the window length. The two channels of the mixture signal are represented by the matrices $X_L(f,t)$ and $X_R(f,t)$, with $f$ being the frequency bin index and $t$ the time frame index.

The signal processing includes the following steps:

1. A pre-processing algorithm reduces the amount of decorrelated energy between the two input channels resulting in $X'_L(f,t)$ and $X'_R(f,t)$, see Sec. 2.1.

2. Multiple separation algorithms are run in parallel, each producing spectral weighting matrices $G_{m,\{L,R\}}(f,t)$ for both input channels $L$ and $R$ to obtain estimates of the dialogue content in them, see Secs. 2.2–2.5.
3. A late fusion stage combines the separation masks from $M$ different modules into one, see Sec. 2.6.
4. A post-processing algorithm is used to smooth the output trading some interferer attenuation performance for a higher perceptual quality, see Sec. 2.7.
5. The final spectral weighting masks $G^{out}_{\{L,R\}}(f,t)$ are applied on the input signal channels for obtaining estimates of the dialogue

$$\hat{X}_{D,\{L,R\}}(f,t) = X'_{\{L,R\}}(f,t) G^{out}_{\{L,R\}}(f,t) \tag{1}$$

and the background

$$\hat{X}_{BG,\{L,R\}}(f,t) = X_{\{L,R\}}(f,t) - \hat{X}_{D,\{L,R\}}(f,t). \tag{2}$$

The processing modules are discussed in more detail in the following. Some of them are single-channel methods; these take the mid signal of the mid/side representation of a stereo signal as the input.

### 2.1 Primary / Ambient Decomposition

The Primary / Ambient Decomposition (PAD) module attenuates the ambient sound components in the input. This can be implemented in various ways, see, e.g., [48–50], with the main principle being the discrimination between direct primary sound sources and diffuse ambience components. The PAD algorithm operates by finding a rotation of the stereo scene that makes the energies of the rotated channels equal, based on the assumption that by doing this the center of the rotated scene points at the primary direct audio source. A center extraction algorithm is used for splitting the signal into primary and ambient components before reversing the rotation.

### 2.2 Center Extraction

The Center Extraction (CE) module relies on the assumption that the dialogue is panned to the center of the stereo scene. This assumption has proven its usability already, e.g., in [30, 32]. The implementation used here is based on the ratio between the smaller and the larger magnitude of the input channels:

$$G_{CE}(f,t) = \frac{\min\left(|X'_L(f,t)|, |X'_R(f,t)|\right)}{\max\left(|X'_L(f,t)|, |X'_R(f,t)|\right)}. \tag{3}$$

The same separation mask is used for both input channels ensuring that the spatial properties of the signal are not distorted.

### 2.3 Non-Negative Matrix Factorization

The single-channel NMF method is derived from a method enforcing block-sparsity [38]. It uses a speech spectral basis dictionary of 75 groups, each with 10 entries, and 20 free bases to model the background. For complexity reasons, the frequency resolution of the signal magnitude spectrum is reduced using a close-to-logarithmic mapping from the 1025 bins to 192 bands before the computation and this is inverted after the source separation mask has been computed. The speech separation mask is obtained from the mixture magnitude spectrogram $X_{in}(f,t)$ and the background magnitude estimate $\hat{X}_{BG}(f,t)$ with

$$G_{NMF}(f,t) = \frac{\max\left(0, X^2_{in}(f,t) - \hat{X}^2_{BG}(f,t)\right)}{X^2_{in}(f,t)}. \tag{4}$$

### 2.4 Harmonic / Percussive / Residual Decomposition

The Harmonic / Percussive / Residual (HPR) decomposition is a single-channel method that applies a median filter to the input magnitude spectrum along either time or frequency axis producing a magnitude spectrum with enhanced sustained or percussive content [51]. Three separation masks are obtained: one for extracting the harmonic components, one for the percussive components, and one for the components not showing a clear harmonic or percussive structure. In [51] the masks are binary, while soft masks are used in our implementation. Speech contains both harmonic (e.g., vowels) and percussive components (e.g., fricatives), so we run HPR with rather extreme parameters compared to [51]. $G_H(f,t)$ and $G_P(f,t)$ are obtained (with values in the range $0-1$), where $G_H(f,t)$ extracts sustained sounds clearly longer than vowels and $G_P(f,t)$ extracts highly dynamic percussive components longer than consonants and being energetic outside the frequency regions characteristic of fricatives (e.g., explosions and shooting). The estimated sustained sounds and explosions are suppressed by the residual separation mask

$$G_{HPR}(f,t) = \min(1 - G_H(f,t), 1 - G_P(f,t)). \tag{5}$$

### 2.5 Single-Channel Speech Enhancement

The Single-Channel Speech Enhancement (SCSE) module applies a recursive averaging algorithm based on speech presence probability. This module makes the assumption that the noise has slow-varying second-order statistics that can be estimated in individual frequency bands as a weighted average of past noise estimates and the present noisy speech power spectrum. The weights change adaptively depending on the speech presence probability, ideally by updating the noise estimate very fast during speech absence and (almost) not updating during speech presence. Our updating rule is based on the likelihood ratio, calculated assuming Gaussian spectral components as in [26]. The noise power estimate is used to compute the weighting mask $G_{SCSE}(f,t)$ as the log-spectral amplitude estimator from Ephraim and Malah [52].

### 2.6 Fusion of Separation Modules

After obtaining the spectral weighting masks for dialogue content separation from multiple modules, these are combined with late fusion. The main task for the fusion is that it should improve the performance compared to the single best module, e.g., by locally selecting the best module based on a quality prediction [53], by using a DNN for combining the separation results [54, 55], or by a weighted

Table 1. Separation performance comparison of the proposed combined method (*All*), and the individual separation modules (Secs. 2.2–2.5). The first three rows are the BSSEval measures (in dB) for the separated dialogue and the last two rows the change from the input mixture. In each cell, the first value is the mean over items for the stereo input data, and the second value is the mean over items for mono input data.

| | All | CE | NMF | HPR | SCSE |
|---|---|---|---|---|---|
| SIR | **14.7 / 12.3** | 9.4 / 2.8 | 10.8 / 10.9 | 3.3 / 1.6 | 6.6 / 7.2 |
| SDR | **8.5 / 8.0** | 7.4 / 2.7 | 7.3 / 7.6 | 2.9 / 1.3 | 5.2 / 5.7 |
| SAR | 10.0 / 10.5 | 13.1 / **20.3** | 0.6 / 11.1 | **16.6** / 16.8 | 12.8 / 13.1 |
| ΔSIR | **13.1 / 9.4** | 7.8 / –0.2 | 9.2 / 7.9 | 1.6 / –1.4 | 5.0 / 4.3 |
| ΔSDR | **6.8 / 5.1** | 5.7 / –0.3 | 5.7 / 4.6 | 1.3 / –1.7 | 3.6 / 2.7 |

average of the separation results after predicting the weighting [56]. After experimenting with various fusion methods, including DNN-based regression [55], the proposed system uses a much simpler approach of an element-wise minimum of the dialogue separation masks:

$$G_{\{L,R\}}(f,t) = \min\left(G_{1,\{L,R\}}(f,t), ..., G_{M,\{L,R\}}(f,t)\right). \tag{6}$$

For each time/frequency-tile the maximum background attenuation from multiple separation methods is selected, assuming the estimation error being mainly undersuppression of the background, e.g., when the stereo modules have no effect on mono input. Compared to the more complex approaches tested, minimum fusion provides better performance while being computationally inexpensive.

### 2.7 Musical Noise Reduction

Musical noise is a recurrent issue for spectral weighting techniques for source separation. Due to localized estimation errors isolated peaks may appear in the processed spectrum, resulting in perceptually annoying wobbling sounds with fast changing pitch. In order to reduce this effect, the adaptive mask smoothing proposed in [57] is adopted as post-filter, applied on $G_{\{L,R\}}(f,t)$, and $G^{out}_{\{L,R\}}(f,t)$ is obtained.

### 2.8 Discussion

Source separation methods in reality are not able to provide ideal separation, but the result still contains both crosstalk and artifacts. The important question is if the result still is *good enough for the intended application*. In an experiment from the BBC, they found that attenuating the background music by 1.4 dB from the default level "allowed many more people to understand what was being said without compromising the editorial vision" [58]. Reproducing a mixture of the separated dialogue and background reduces the prominence and audibility of source separation artifacts and this fits well together with the idea of DE application that the mixing ratio between the dialogue and background is only *adjusted*. Since the actual adjustment of the mixing ratio is done by the end-user, he can decide it depending on the personal needs and opinion of the quality within the limits set by the broadcaster.

## 3 EVALUATION

### 3.1 Evaluation of Fusion

First we evaluate the performance of each separation module from Secs. 2.2–2.5 independently (including the pre- and post-processing stages, Secs. 2.1, 2.7) and compare it with the result after the fusion (Sec. 2.6). The data consists of 11 synthetic items, each 10 s in length, with mono centered dialogue and a stereo background. The separation performance is evaluated using the BSSEval [59] measures of Signal-to-Interferer Ratio (SIR), Signal-to-Distortion Ratio (SDR), and Signal-to-Artifacts Ratio (SAR), and the change of SIR (ΔSIR) and SDR (ΔSDR) from the input mixture. The proposed system is also tested using mono inputs obtained by downmixing the stereo items. The evaluation results are given in Table 1, from which we see that the fusion result outperforms the individual modules for both mono and stereo inputs.

### 3.2 Subjective Evaluation in Application

**Methodology.** In [10] we presented the Adjustment/Satisfaction Test (A/ST) where the participants interact with a user-adjustable system and their adjustments and the resulting satisfaction levels are studied. This allows analyzing to what extent the available personalization is used and quantifying the quality of experience improvement. We use the A/ST here in the form of a DE application, in which the user adjusts the dialogue level in the signal, comparing the proposed *blind source separation* (BSS) system, referred to as $S_{BSS}$, with using the *original objects* (OO), referred to as $S_{OO}$.

**Environment.** The experiment is carried out in a listening room resembling a quiet low-reverberant living room with a 5.1 setup with high-end studio monitors positioned according to [60]. The user interface is displayed on a TV positioned above the center loudspeaker. The participants sit on a chair with fixed position, and they can control the relative level of the dialogue via a rotating knob.

**Participants.** The test involves 14 German participants with good knowledge of the English language. They have taken part in a sensory test before; they have normal hearing; and they are voluntary, remunerated, and between 22 and 38 years old (median age is 25).

**Instructions.** The participants are not informed that two different systems are tested. Still, the test instructions

Table 2. Initial loudness difference for the center channel ($LD_0^c$) given in Loudness Units (LU), percentage of frames with active speech (AS%), and type of background for each item.

| Item | $LD_0^c$ | AS% | Background |
|---|---|---|---|
| 1Docu | 7.18 | 78 | Ambient music |
| 2VOWar | 3.64 | 67 | War shooting and explosions |
| 3Cheer | 18.6 | 90 | Cheering crowd indoors |
| 4Docu | 7.50 | 67 | Ambient music |
| 5Docu | 14.9 | 51 | Instrumental classical music |
| 6Docu | 8.20 | 82 | Instrumental jazz music and ambient noise |
| 7Cheer | 5.49 | 87 | Cheering crowd outdoors |
| Median | 7.50 | 78 | |

Table 3. ANOVA of the LD adjustments: degrees of freedom (d.f.), effect size $\eta^2$ (as a percentage of the total variation explained), and p-values (if lower than 0.05, we reject the null hypothesis).

| Effect | d.f. | $\eta^2$ (%) | p |
|---|---|---|---|
| Subject | 13 | 21.7 | 0.02 |
| $S_{OO}$ / $S_{BSS}$ | 1 | 15.8 | 0.00 |
| Music Background | 1 | 11.2 | 0.00 |
| Item | 5 | 7.8 | 0.00 |
| Item × Subject | 65 | 15.3 | 0.03 |
| Item × $S_{OO}$ / $S_{BSS}$ | 5 | 6.6 | 0.00 |
| Subject × Music Back. | 13 | 6.4 | 0.03 |
| $S_{OO}$ / $S_{BSS}$ × Music Back. | 1 | 1.1 | 0.01 |
| $S_{OO}$ / $S_{BSS}$ × Subject | 13 | 2.0 | 0.46 |
| Error | 78 | 12.0 | |

mention the possibility that the personalization might introduce quality degradations (even if this is true only for $S_{BSS}$). The participants are asked to adjust the relative dialogue level in the the audio so that they can easily follow the text, yet keeping the background enjoyable, i.e., *to find the best compromise between the preferred relative dialogue level and a sound quality they would accept* in television. After the adjustment, the participants are asked to rate the difference in satisfaction between the initial mix and the personalized version. The satisfaction is assessed directly after adjusting each item, i.e., using the "Experience configuration" from [10].

**Stimulus.** Seven 5.1 audio excerpts for TV with the length of 7–12 s and the sampling frequency of 48 kHz are considered as test material: four excerpts feature female English talkers and three feature male English talkers. These excerpts are selected to have loud background, potentially making the dialogue tiring to follow. Each item is presented once with $S_{OO}$ and once with $S_{BSS}$ in a pseudo-random order. The repetitions of an item are not directly consecutive but interleaved with other items.

Table 2 shows the initial loudness differences (LDs)[1] between the dialogue and the background in the center channel ($LD_0^c$). The initial LDs considering all the 5.1 channels are shown by the dashed black line in Fig. 2 (*initial* $LD_0$). The listeners are able to modify the relative level of the speech from –10 to +20 LU with respect to the initial $LD_0$ while operating $S_{OO}$. While operating $S_{BSS}$ the available range depends on the separation performance and is item-dependent. The maximum available LD is shown by the dashed blue line in in Fig. 2 (*max* $S_{BSS}$). All the items are loudness-normalized to have equal integrated loudness [61] both in their initial and adjusted versions.

**Results.** Fig. 2 depicts the mean of the listeners' adjustments and satisfaction levels (solid lines) together with box plots[2]. High subjective variance is visible, i.e., subjects have very different preferences for the relative level of the dialogue. This confirms that a unique *one-size-fits-all* mix would hardly satisfy all listeners and DE is desired.

This is also supported by the ANOVA of the LD adjustments, where four factors are considered: item, subject, type of used system ($S_{OO}$ / $S_{BSS}$), and type of background (i.e., if the background contains music or not, referred to as "Music Background"). Subject is considered as random factor, as it consists of samples randomly taken from the relevant population on which we would like to generalize. Item is nested inside Music Background. Table 3 reports the ANOVA results. The factor "Subject" accounts alone for 21.7% of the total variation.

Fig. 2 shows also that slightly lower levels of adjustment are preferred for $S_{BSS}$ than for $S_{OO}$, resulting in lower satisfaction. In fact, the subjects have to *trade-off between the desired LD* (selected while operating $S_{OO}$) *and the distortions*, which $S_{BSS}$ introduces for higher levels of adjustment. Yet, the difference between $S_{BSS}$ and $S_{OO}$ in terms of selected LD is smaller than 1 LU on average and both systems clearly increase the satisfaction. Averaging over the listeners, the satisfaction increase correlates with the LD adjustment (i.e., the difference between the selected LD and the initial LD) with Pearson's $r = 0.9$.

To support this observation, an ANOVA is run on the satisfaction scores considering LD adjustment, item, subject, and music background as main factors and no interaction is considered. All four factors result in statistical significance ($p < 0.05$), with LD adjustment accounting alone for 62% of the total variation, subject accounting for 12%, and the remaining factors together accounting for 6%. It can be concluded that the adjustment has a *noticeable and positive effect* and the personalization offered by $S_{BSS}$ is desired, despite the distortions potentially introduced.

These observations confirm the results obtained in [10], where the same system $S_{BSS}$ was tested using stereo material with speech panned to the center, i.e., $S_{BSS}$ could exploit also stereo methods such as PAD and CE. Fig. 3 summarizes the results from this previous evaluation.

[1] Loudness is herein meant as integrated loudness as per BS.1770-4 [61] and measured in Loudness Units (LU).

[2] The boxes correspond to the 25/75% quantiles of the data, the central black bar corresponds to the median, the whiskers indicate the minimum or maximum points within 1.5 IQR (interquartile range, and points are displayed with a cross if they are within 1.5–3 times the IQR and with a circle if they exceed 3 IQR.

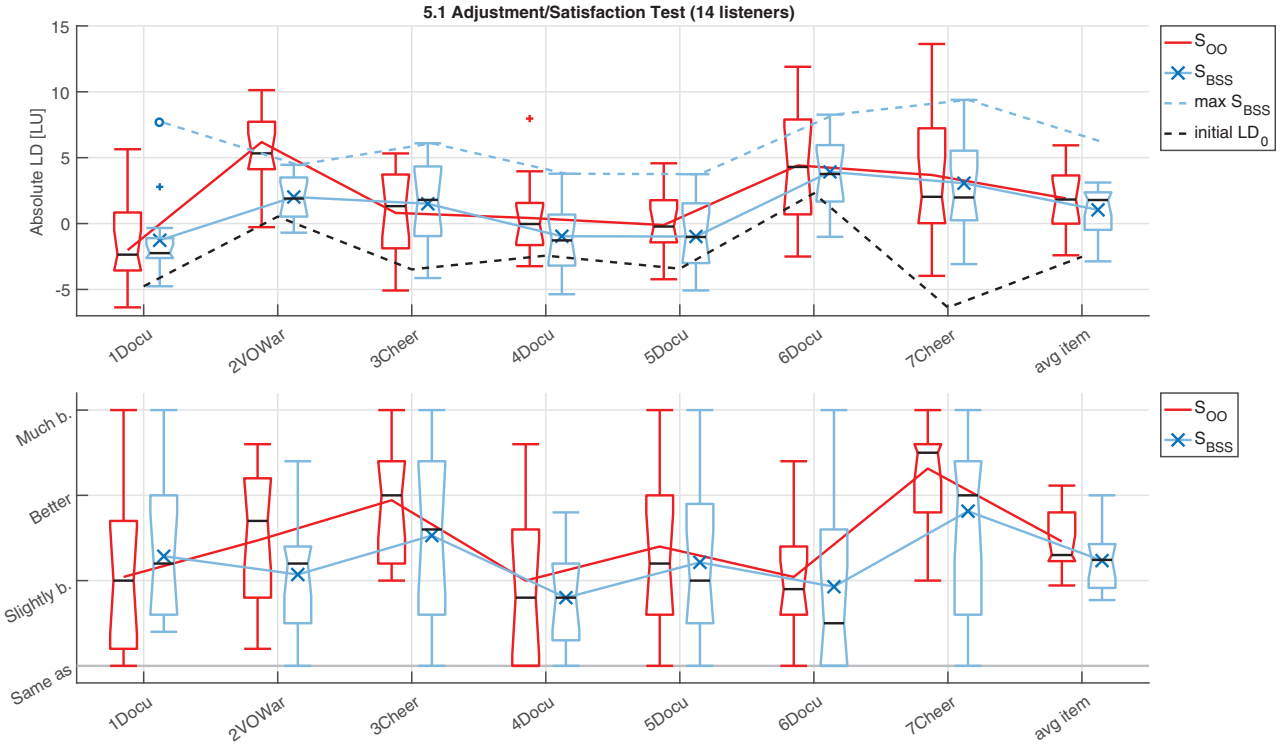


Fig. 2. Mean selections and box plots for the preferred LD (upper plot) and resulting satisfaction levels (lower plot). $S_{OO}$ (main red lines) is compared with $S_{BSS}$ (main blue lines with crosses). Test signals are 5.1 with dialogue mixed in the center channel.

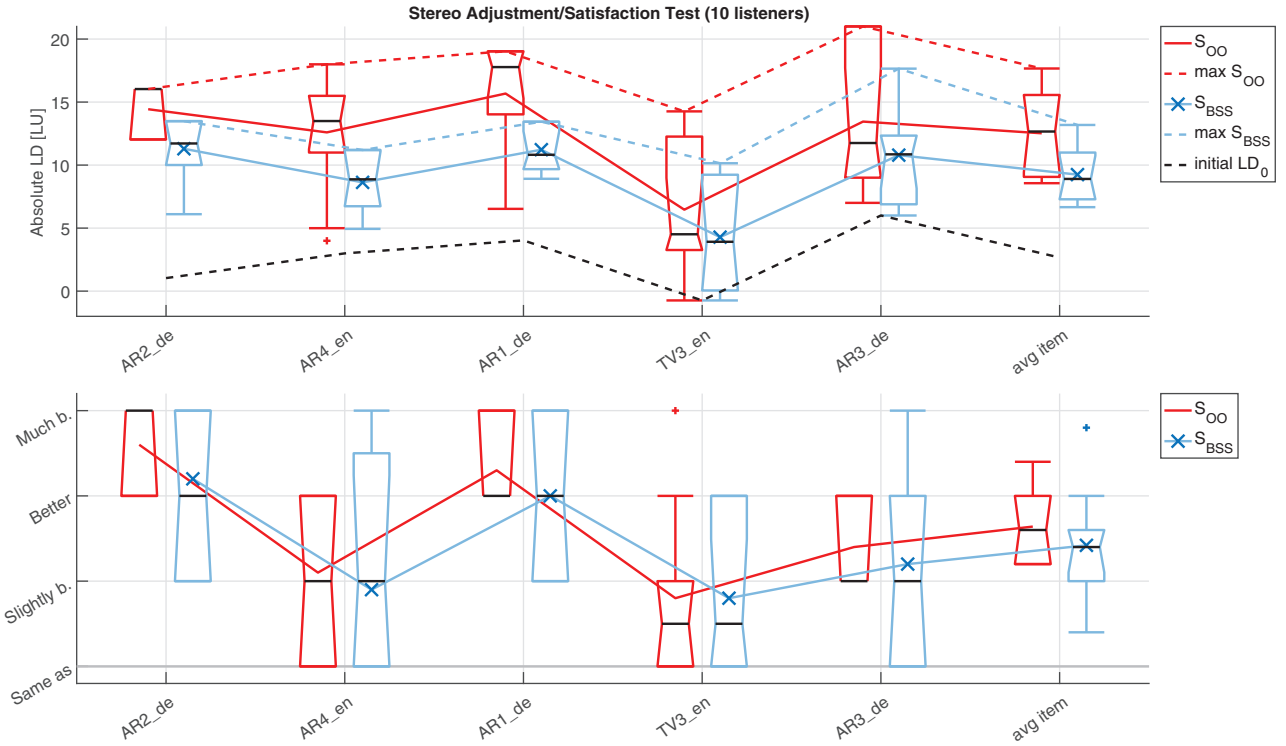


Fig. 3. Previous evaluation with stereo material and speech panned to the center ©2018 IEEE. Reprinted, with permission, from [10].

## 4 CONCLUSIONS

Dialogue Enhancement (DE) enables the delivery of optimal dialogue mixing to every listener, be it in terms of intelligibility or for aesthetic preference. This paper investigated technology for the implementation of DE in object-oriented broadcasting, such as MPEG-H. A special focus of the paper was on the use of source separation methods to extract dialogue and background from the complex sound mixture also in the case when these have not been made available during the production process, i.e., for legacy content. The presented source separation technology integrates several separation approaches with known limitations into a more powerful overall architecture.

The second main focus of the paper was on the evaluation of the subjective benefit of individually adjustable DE using the Adjustment/Satisfaction Test. The listeners made extensive use of the dialogue level personalization and the preferred dialogue level had a high variance among the listeners indicating the need for this functionality. The use of the personal adjustment increased the listener satisfaction clearly. The extensive use of personalization and increased satisfaction were observed also when using the proposed source separation method for obtaining the dialogue and background objects. The cost of the imperfect source separation compared to using the original objects is visible in the user satisfaction as a slightly smaller improvement.

In summary, it was shown that the benefits of object-based audio, as they are used in modern broadcasting systems, can also be used when broadcasting legacy content that was not produced in an object-oriented way by using current source separation technology. This may lower the transition barrier for the adoption of object-oriented broadcasting standards.

In the future, substantial further improvements in source separation can be expected inspired from the field of deep learning methods, both for the fusion of individual source separation module outputs and for the source separation task itself.

## 5 REFERENCES

[1] M. Armstrong, "From Clean Audio to Object Based Broadcasting," *Research & Development White Paper WHP324*, BBC (2016 Sep.).

[2] B. Shirley, et al., "Personalized Object-Based Audio for Hearing Impaired TV Viewers," *J. Audio Eng. Soc.*, vol. 65, pp. 293–303 (2017 Apr.), https://doi.org/10.17743/jaes.2017.0005.

[3] T. Walton, et al., "Does Environmental Noise Influence Preference of Background-Foreground Audio Balance?" presented at the *141st Convention of the Audio Engineering Society* (2016 Sep.), convention paper 9637.

[4] P. Mapp, "Intelligibility of Cinema & TV Sound Dialogue," presented at the *141st Convention of the Audio Engineering Society* (2016 Sep.), convention paper 9632.

[5] M. Florentine, "Speech Perception in Noise by Fluent, Non-Native Listeners," *J. Acoust. Soc. Amer.*, vol. 77, no. S107 (1985), https://doi.org/10.1121/1.2022152.

[6] A. Warzybok, et al., "Influence of the Linguistic Complexity in Relation to Speech Material on Non-Native Speech Perception in Noise," *Fortschritte der Akustik - DAGA'2010*, pp. 987–988 (2010 Mar.).

[7] C. G. Clopper and A. R. Bradlow, "Perception of Dialect Variation in Noise: Intelligibility and Classification," *Language and Speech*, vol. 51, no. 3, pp. 175–198 (2008 Sep.), https://doi.org/10.1177/0023830908098539.

[8] P. Adank, et al., "Comprehension of Familiar and Unfamiliar Native Accents Under Adverse Listening Conditions," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 35, no. 2, pp. 520–529 (2009), https://doi.org/10.1037/a0013552.

[9] H. Fuchs and D. Oetting, "Advanced Clean Audio Solution: Dialogue Enhancement," *SMPTE Motion Imaging J.*, vol. 123, no. 5, pp. 23–27 (2014 Jul.), https://doi.org/10.5594/j18429.

[10] M. Torcoli, et al., "The Adjustment / Satisfaction Test (A/ST) for the Evaluation of Personalization in Broadcast Services and its Application to Dialogue Enhancement," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 524–538 (2018 Jun.), https://doi.org/10.1109/TBC.2018.2832458.

[11] B. G. Shirley and P. Kendrick, "ITC Clean Audio Project," presented at the *116th Convention of the Audio Engineering Society* (2004 May), convention paper 6027.

[12] B. Shirley and R. Oldfield, "Clean Audio for TV Broadcast: An Object-Based Approach for Hearing-Impaired Viewers," *J. Audio Eng. Soc.*, vol. 63, pp. 245–256 (2015 Apr.), https://doi.org/10.17743/jaes.2015.0017.

[13] J. Paulus, et al., "MPEG-D Spatial Audio Object Coding for Dialogue Enhancement (SAOC-DE)," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9220.

[14] J. Herre, et al., "MPEG-H Audio—The New Standard for Universal Spatial / 3D Audio Coding," *J. Audio Eng. Soc.*, vol. 62, pp. 821–830 (2014 Dec.), https://doi.org/10.17743/jaes.2014.0049.

[15] ISO/IEC, "Information technology—High efficiency coding and media delivery in heterogeneous environments—Part 3: 3D audio," *International Standard 23008-3* (2015).

[16] ISO/IEC, "Information technology—High efficiency coding and media delivery in heterogeneous environments—Part 3: 3D audio, AMENDMENT 3," *MPEG-H 3D Audio Phase* 2 (2015).

[17] ISO/IEC, "Information technology—High efficiency coding and media delivery in heterogeneous environments—Part 3: 3D audio," *International Standard 23008-3:2018*, Second Ed. (2018).

[18] "EBU and Fraunhofer IIS Conducted Live MPEG-H Audio Production Trial at Eurovision Song Contest 2018," http://www.audioblog.iis.fraunhofer.com/ebu-fraunhofer-mpegh-eurovision/, accessed: 2018-06-26.

[19] "Successful Terrestrial and Satellite Reception of MPEG-H Audio During the Roland Garros French Open," http://idfrancetv.fr/successful-terrestrial-and-satellite-reception-of-mpeg-h-audio-during-the-roland-garros-french-open/, accessed: 2018-06-26.

[20] R. L. Bleidt, et al., "Development of the MPEG-H TV Audio System for ATSC 3.0," *IEEE Trans. Broadcast.*, vol. 63, no. 1, pp. 202–236 (2017 Mar.), https://doi.org/10.1109/TBC.2017.2661258.

[21] ETSI TS 101 154 v2.4.1, "Digital Video Broadcasting (DVB); Specification for the Use of Video and Audio Coding in Broadcast and Broadband Applications" (2018).

[22] ETSI TS 103 285 v1.2.1, "Digital Video Broadcasting (DVB); MPEG-DASH Profile for Transport of ISO BMFF Based DVB Services over IP Based Networks" (2018).

[23] Telecommunications Technology Association of Korea, "Transmission and Reception for Terrestrial UHDTV Broadcasting Service, Rev. 1," KO-07.0127R1 (2016).

[24] R. Martin, "Spectral Subtraction Based on Minimum Statistics," *7th Eur. Signal Process. Conf.*, pp. 1182–1185 (1994 Sep.).

[25] G. Doblinger, "Computationally Efficient Speech Enhancement by Spectral Minima Tracking In Subbands," *Eurospeech*, pp. 1513–1516 (1995).

[26] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum-Mean Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 6, pp. 1109–1121 (1984 Dec.), https://doi.org/10.1109/TASSP.1984.1164453.

[27] L. Lin, et al., "Adaptive Noise Estimation Algorithm for Speech Enhancement," *Electronic Letters*, vol. 39, no. 9, pp. 754–755 (2003 May), https://doi.org/10.1049/el:20030480.

[28] H. Hirsch, and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," *IEEE Int. Conf. Acoust. Speech Signal Process*., pp. 153–156 (1995 May), https://doi.org/10.1109/ICASSP.1995.479387.

[29] P. C. Loizou, *Speech Enhancement —Theory and Practice* (CRC Press, Boca Raton, FL, USA, 2007).

[30] E. Vickers, "Two-to-Three Channel Upmix for Center Channel Derivation and Speech Enhancement," presented at the *127th Convention of the Audio Engineering Society* (2009 Oct.), convention paper 7917.

[31] J. T. Geiger, et al., "Dialogue Enhancement of Stereo Sound," *23rd Eur. Signal Process. Conf.*, pp. 874–878 (2015 Aug.), https://doi.org/10.1109/EUSIPCO.2015.7362507.

[32] A. Craciun, et al., "An Evaluation of Stereo Speech Enhancement Methods for Different Audio-Visual Scenarios," *23rd Eur. Signal Process. Conf.*, pp. 2048–2052 (2015 Aug.), https://doi.org/10.1109/EUSIPCO.2015.7362744.

[33] D. Barry, et al., "Sound Source Separation: Azimuth Discrimination and Resynthesis," presented at the *Int. Conf. Digital Audio Effects* (2004 Oct.).

[34] A. Jourjine, et al., "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures," *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 2985–2988 (2000 Jun.), https://doi.org/10.1109/ICASSP.2000.861162.

[35] D. D. Lee and H. S. Seung, "Algorithms for Non-Negative Matrix Factorization," in T. Leen, T. Dietterich, and V. Tresp (Eds.), *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562 (MIT Press, 2001).

[36] F. Weninger, et al., "Discriminative NMF and its Application to Single-Channel Source Separation," *Interspeech 2014*, pp. 865–869 (2014 Sep.).

[37] D. L. Sun and G. J. Mysore, "Universal Speech Models for Speaker Independent Single Channel Source Separation," *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 141–145 (2013 May), https://doi.org/10.1109/ICASSP.2013.6637625.

[38] M. Kim and P. Smaragdis, "Mixtures of Local Dictionaries for Unsupervised Speech Enhancement," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 293–297 (2015 Mar.), https://doi.org/10.1109/LSP.2014.2346506.

[39] J. Tchorz and B. Kollmeier, "SNR Estimation Based on Amplitude Modulation Analysis with Applications to Noise Suppression," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 184–192 (2003 May), https://doi.org/10.1109/TSA.2003.811542.

[40] M. Kleinschmidt and V. Hohmann, "Sub-Band SNR Estimation Using Auditory Feature Processing," *Speech Communication*, vol. 39, pp. 47–64 (2003), https://doi.org/10.1016/S0167-6393(02)00058-4.

[41] C. Uhle, et al., "Speech Enhancement of Movie Sound," presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), convention paper 7628.

[42] Y. Jiang, et al., "Binaural Classification for Reverberant Speech Segregation Using Deep Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12 (2014 Dec.), https://doi.org/10.1109/TASLP.2014.2361023.

[43] S. Araki, et al., "Exploring Multi-Channel Features for Denoising-Autoencoder-Based Speech Enhancement," presented at the *IEEE Int. Conf. Acoust. Speech Signal Process.* (2015 Apr.), https://doi.org/10.1109/ICASSP.2015.7177943.

[44] N. Fan, J. Du, and L.-R. Dai, "A Regression Approach to Binaural Speech Segregation via Deep Neural Networks," presented at the *10th Int. Symp. Chin. Spoken Lang. Process.* (2016 Oct.), https://doi.org/10.1109/ISCSLP.2016.7918387.

[45] J. M. Martín-Doñas, et al., "Dual-Channel DNN-Based Speech Enhancement for Smartphones," presented at the *IEEE 19th Int. Workshop Multimedia Signal Process.* (2017 Oct.), https://doi.org/10.1109/MMSP.2017.8122273.

[46] X. Zhang and D. Wang, "Deep Learning Based Binaural Speech Separation in Reverberant Environments," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 5 (2017 May), https://doi.org/10.1109/TASLP.2017.2687104.

[47] D. Wan and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726 (2018 Oct.), https://doi.org/10.1109/TASLP.2018.2842159.

[48] C. Avendano and J.-M. Jot, "A Frequency-Domain Approach to Multichannel Upmix," *J. Audio Eng. Soc.*, vol. 52, pp. 740–749 (2004 Jul./Aug.).

[49] J. Merimaa, et al., "Correlation-Based Ambience Extraction from Stereo Recordings," presented at the *123rd Convention of the Audio Engineering Society* (2007 Oct.), convention paper 7282.

[50] K. M. Ibrahim and M. Allam, "Primary-Ambient Source Separation for Upmixing to Surround Sound Systems," *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 431–435 (2018 Apr.), https://doi.org/10.1109/ICASSP.2018.8461459.

[51] J. Driedger, et al., "Extending Harmonic-Percussive Separation of Audio Signals," *15th Int. Soc. Music Inform. Retrieval Conf.*, pp. 611–616 (2014 Oct.).

[52] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans Acoust. Speech Signal Process.*, vol. 33, no. 2, pp. 443–445 (1985 Apr.), https://doi.org/10.1109/TASSP.1985.1164550.

[53] E. Manilow, et al., "Predicting Algorithm Efficacy for Adaptive Multi-Cue Source Separation," *2017 IEEE Workshop Appl. of Signal Process. Audio and Acoustics*, pp. 274–278 (2017 Oct.), https://doi.org/10.1109/WASPAA.2017.8170038.

[54] E. M. Grais, et al., "Combining Mask Estimates for Single Channel Audio Source Separation Using Deep Neural Networks," *Interspeech 2016*, pp. 3339–3343 (2016 Sep.), https://doi.org/10.21437/Interspeech.2016-216.

[55] A. Ragano and A. Hines, "Exploring a Perceptually-Weighted DNN-Based Fusion Model for Speech Separation," *26th AIAI Irish Conf. Artif. Intell. and Cogn. Sci.*, pp. 21–32 (2018 Dec.).

[56] X. Jaureguiberry, et al., "Fusion Methods for Speech Enhancement and Audio Source Separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 7, pp. 1266–1279 (2016 Jul.), https://doi.org/10.1109/TASLP.2016.2553441.

[57] T. Esch and P. Vary, "Efficient Musical Noise Suppression for Speech Enhancement System," *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 4409–4412 (2009 Apr.), https://doi.org/10.1109/ICASSP.2009.4960607.

[58] "Is the Background Music Too Loud?" http://www.bbc.co.uk/blogs/tv/2011/03/is-the-background-music-too-loud. shtml, accessed: 2018-06-26.

[59] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469 (2006 Jun.), https://doi.org/10.1109/TSA.2005.858005.

[60] ITU-R Recommendation BS.775-3, "Multichannel stereophonic sound system with and without accompanying picture" (2012 Aug.).

[61] ITU-R Recommendation BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level" (2015 Oct.).

## THE AUTHORS


Jouni Paulus


Matteo Torcoli


Christian Uhle


Jürgen Herre


Sascha Disch


Harald Fuchs

Jouni Paulus received the M.Sc.(Eng.) and D.Sc.(Tech.) degrees in information technology from Tampere University of Technology (TUT), in 2002 and 2010, respectively. From 2002 to 2010 he was working as a researcher at the Department of Signal Processing at TUT with the topic of signal-based music content analysis. In 2010 he joined Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany, as a research scientist, and as a member of the International Audio Laboratories Erlangen. Dr. Paulus has contributed to the development and standardization of MPEG-D SAOC and MPEG-H 3D Audio. His current research interests as a senior scientist at Fraunhofer IIS include object-based and spatial audio coding, informed and blind source separation, machine learning for audio applications, speech intelligibility enhancement, and subjective evaluation of the resulting audio processing algorithms.

•

Matteo Torcoli received his B.Sc. degree in computer engineering from the University of Brescia in 2011 and his M.Sc. degree in sound and music computer engineering from the Politecnico di Milano in 2014, cum laude. He worked on his M.Sc. thesis on dereverberation for next-generation hearing aids at the International Audio Laboratories Erlangen. He then joined the Audio and Media Technologies division of Fraunhofer IIS, where he is currently working as R&D engineer. His research focus is on applying digital signal processing and machine learning for developing accessibility features. In particular, he has been working on dialogue enhancement, on ways to enable it also without the original audio objects, and on the subjective and objective evaluation of such an experience.

•

Christian Uhle is chief scientist in the Audio and Media Technologies division of the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany. He received the diploma engineer degree and the Ph.D. degree in electrical engineering from the Technical University of Ilmenau, Germany, in 1997 and 2008, respectively. From 1998 to 2000 he developed a real-time operation system for digital signal processors at the Technical University of Ilmenau. From 2000 until 2005 he was research associate and doctoral student at the Fraunhofer Institute for Digital Media Technology (IDMT) working on the semantic analysis of musical audio signals. Since 2006 he is research associate at Fraunhofer IIS. His research activities comprise semantic audio processing, blind source separation, dialog enhancement, digital audio effects, automotive sound reproduction, and natural language processing. Dr. Uhle is a member of the AES and chairs the AES Technical Committee on Semantic Audio Analysis.

•

Jürgen Herre joined the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany, in 1989. Since then he has been involved in the development of perceptual coding algorithms for high quality audio, including the well-known ISO/MPEG-Audio Layer III coder (aka "MP3"). In 1995, Dr. Herre joined Bell Laboratories for a PostDoc term working on the development of MPEG-2 Advanced Audio Coding (AAC). By the end of 1996 he went back to Fraunhofer to work on the development of more advanced multimedia technology including MPEG-4, MPEG-7, MPEG-D, and MPEG-H, currently as the Chief Executive Scientist for the Audio & Media Technologies activities at Fraunhofer IIS, Erlangen. In September 2010, Dr. Herre was appointed professor at the University of Erlangen and the International Audio Laboratories Erlangen. Dr. Herre is a fellow of the Audio Engineering Society, co-chair of the AES Technical Committee on Coding of Audio Signals, and vice chair of the AES Technical Council. He served as a member of the IEEE Technical Committee on Audio and Acoustic Signal Processing and as an associate editor of the *IEEE Transactions on Speech and Audio Processing* and is an active member of the MPEG audio subgroup.

•

Sascha Disch received his Dipl.-Ing. degree in electrical engineering from the Technical University Hamburg-Harburg (TUHH) in 1999 and joined the Fraunhofer Institute for Integrated Circuits (IIS) the same year. Ever since he has been working in research and development

of perceptual audio coding and audio processing. From 2007 to 2010 he was a researcher at the Laboratory of Information Technology, Leibniz University Hannover (LUH), receiving his Doctoral Degree (Dr.-Ing.) in 2011. He contributed to the standardization of MPEG Surround, MPEG Unified Speech and Audio Coding (USAC), MPEG-H 3D Audio, and the 3GPP Enhanced Voice Services (EVS) codec. His research interests as a Chief Scientist at Fraunhofer IIS and a member of the International Audio Laboratories Erlangen include waveform and parametric audio coding, audio bandwidth extension, and digital audio effects.

•

Harald Fuchs received his diploma in electrical engineering from the University of Erlangen, Germany, in 1997 and joined Fraunhofer IIS in the same year. From 1997 to 2002 he was a software developer for video codecs and multimedia streaming systems. From 2002 onwards, he concentrated on media system aspects and standardization, contributing to several standardization organizations, including MPEG, DVB, ATSC, DLNA, OMA, OIPF, ISMA, and HbbTV. Since 2011 his main interest is on object-based audio, especially focusing on how media applications can benefit from object-based and next generation audio. As a Senior Business Development Manager, Audio for TV Broadcast, he has taken, specifically, care of enabling MPEG-H Audio in broadcast and streaming systems, like ATSC 3.0 and DVB. From 2013 to 2018 he was group manager semantic audio coding, with the main target of enabling object-based audio for dialogue enhancement and better speech intelligibility in broadcast applications. Since 2017 he is product manager for MPEG-H Audio, and since 2018 he is head of the Media Systems and Applications department.