The LIPS Vision

Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag

The Project

As the third biggest industry sector in Germany, the culture & creative industry is a main driver for stable revenue, and high-quality user experiences, either for live or recorded content. Reaching a high level of excellence requires focusing on innovative ideas and involving diverse technologies that must be implemented and maintained on both sides: the production and remote sides. The technological cornerstone of content production is the PMSE industry. The term Programme Making and Special Events (PMSE) covers all wireless applications in production and event technologies such as wireless cameras, wireless microphones, effects control and so on.

The LIPS (Live Interactive PMSE Services) project works on different services which combine several local PMSE services into one professional production network. The project is co-funded by the German Federal Ministry of Economics and Technology. The overall project budget is about €6.5mio including €3.8mio funding. It started in April 2018 and will finish in September 2020. The project consortium consists of:

- Sennheiser electronic GmbH & Co. KG (SEN)
- Arnold & Richter Cine Technik GmbH & Co. Betriebs KG (ARRI)
- TVN Mobile Production GmbH (TVN)
- Smart Mobile Labs AG (SML)
- Fraunhofer Heinrich Hertz Institute (HHI)
- Friedrich-Alexander-University Erlangen-Nürnberg (FAU)
- Leibniz University Hannover (LUH)

The project targets a highly interactive and immersive linking of spatially separated locations in order to enable future networked activities such as networked music making, remote Audio/Video production, and immersive conferencing. Additionally, LIPS aims to connect cultural events, which take place in the city (e.g. stage performances, operas and philharmonic concerts, talks and political debates, etc.) with participants from rural areas through technological convergence. Based on newly developed IP-based audio/video production equipment, the LIPS project identifies solutions to improve access to and participation in events of the culture and creative industry for a wide range of users and to demonstrate novel smart services in a realistic scenario.

The Use Case

The LIPS Project focuses on two main areas: (1) improving the immersive audio/visual experience for the connected users, as well as (2) developing the technological background for connecting and merging various devices and locations into one production network. The LIPS use-case reflects this goal via two separated but interconnected rooms, enabling the immersive reproduction of the respective other room to support remote rehearsals, remote Audio/Video production and so on.

Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag

The LIPS Vision

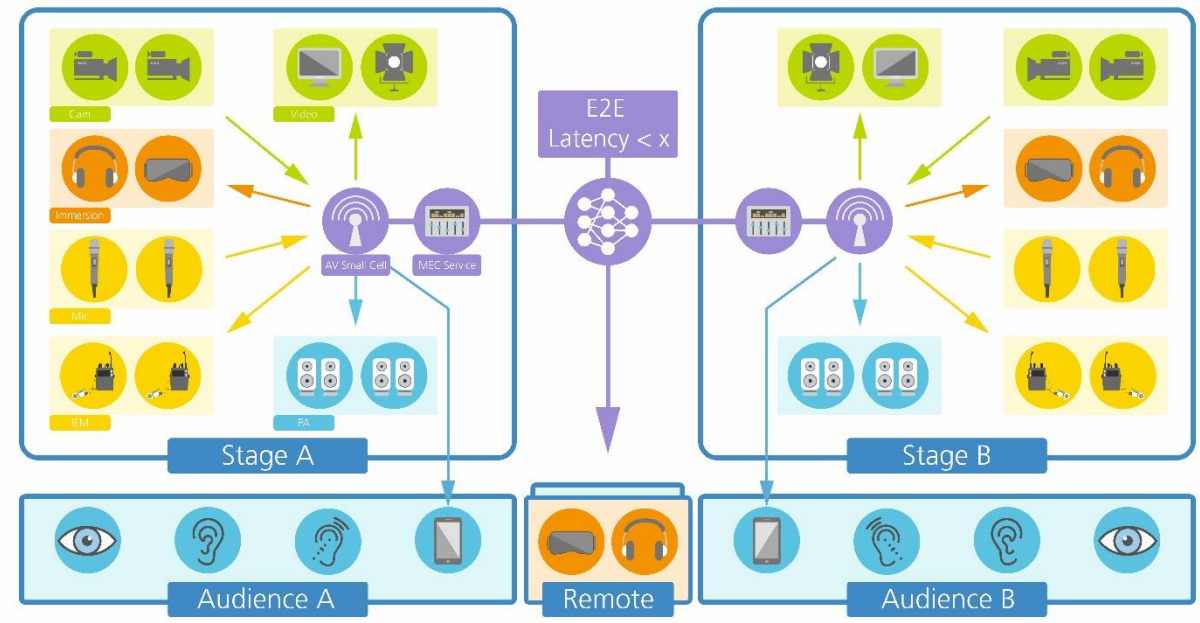

Figure 1: Graphical schematic of the LIPS scenario

Such goals face challenges from a variety of perspectives, not only technological, but also organizational, regulative and finally the acceptance of the users. The technological requirements are well-known and can be specified easily when it comes to latency, jitter, reliability, synchronicity, transmission quality, and data rates (cf. *The Annex*). These metrics are monitored during operation, especially when using commodity IP-based network technologies and must be guaranteed to not exceed a predefined threshold. During a (live) production, the technical equipment must operate error free to enable further processing and distribution. In addition, the acceptance of a new technological change or advancement is increased if it meets already established quality levels or even improves it while simplifying its operation.

The Immersion

As described above, the LIPS Project pursues the idea of realistic joint performances of musicians at different geographic locations. For a natural interaction, an immersive, interactive audio-visual connection between the two locations must be created.

Immersion describes the effect of "diving into" an impressively presented or (virtually) supplemented reality that is perceived as completely real. Audio as well as the created optical ambience are essential elements of the impression of immersion. LIPS considers immersion for the audience, for the artists, and for the remote producer. Furthermore, LIPS looks at interactivity, i.e. the interaction between the users at both locations, who can perceive each other and adapt their behavior.

One approach covered within the LIPS Project is to create a "virtual window" between two linked rooms, which allows a visually and audibly plausible "look" into the connected room (cf. *The Technologies*).

Immersive visual presence places high demands on imaging as well as lighting. Light emission, reflections in the scene, and image capture represent a tightly coupled physical system that needs to be considered together. Human visual perception of images presented on electronic displays is dependent on local ambient light. A defined and matched lighting environment at both locations is

needed to provide a consistent look of image and reality. On the other hand, the natural environment and natural lighting at each side shall be preserved widely to avoid a “studio feeling”. Artificial light sources shall be used only to the extent needed for matching both locations.

Remote objects need to be perceived consistent with local objects in color, brightness and size. The project aims to establish a linear unscaled image path from image acquisition to image reproduction. High contrast ratio and wide color gamut of the image path are needed to deal with the wide range of natural environments.

The Quality

In order to plan, create and maintain a live immersive experience many different parameters of the system must be considered. Latency of the end-to-end transmission is the most critical parameter. Real-time interaction within a live environment can only be experienced when the communication system runtime and its delay is very low. A real world example is, when calling someone using a high latency connection, fluent conversation is very difficult to achieve.

Especially, high latency has a negative effect on music performance. If, for example, two musicians hear their own instrument with very low delay, but the other musician's instrument has a much higher delayed, this leads to irritation and the rhythmic interplay is impaired. This usually leads to a successive deceleration of the playing tempo. A latency of approx. 25 ms [1] (for the one-way transmission path) is still perceived by musicians as "not disturbing" without them having to resort to compensating strategies that reduce the effects of larger latencies.

The second important parameter for audio/video streaming applications is synchronicity. To avoid noticeable gaps in the audio/video playback and to enable further data processing, all involved devices in the processing chain should be synchronized to a global master clock. Synchronicity is important for all parties of an event like the musicians, the production and the distribution. It is essential to know the exact timestamp, a sound or a video frame was recorded, in order to create the intended mix and to have synchronized video and audio streams. In addition, future applications such as volumetric audio/video capture need a very high degree of synchronous operation within the whole production network to enable mixed reality applications.

These two parameters can be measured and summed up in one figure of merit, which is called quality of service (QoS). This QoS is defined as followed in the ITU-T P1.10 [2]:

“The totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service.”

The technical figure of merit defines the quality of the whole system. In contrast to this, quality of experience (QoE) relates to the experience of the user. The ITU defines, that QoE is defined as “*the degree of delight or annoyance of the user of an application or service*” [2], [3]. Hence, the expectation of a user about the application is also very important as it influences the QoE.

However, QoE cannot exist without QoS. This becomes more obvious, if we look back at the LIPS Use Case with two musicians at two different sites. If the latency of the connection gets too high, the musicians are not able to play together anymore, because the playing tempo will decelerate. Thus, the QoE for the musicians as well as for the audience will suffer.

But this is not the whole truth. If there is an appropriate countermeasure, the QoE for the users has not to suffer even when the QoS declines. In the LIPS environment, an example for such a

countermeasure is the global metronome. This metronome creates a synchronous beat at both sides. Even when the latencies between the two sites rises, the musicians are able to play synchronously and the live performance might not suffer. This leads to constant high QoE for the audience of the live concert. However, the compensation of the metronome is also limited to certain latencies, depending on the music genre and level of expertise [4].

The Technologies

The Audio

For creating a plausible and immersive audio experience for the musicians at both sites, modern techniques of spatial audio capturing and reproduction are available.

On the capturing side special microphone arrays, e.g. spherical arrays, capture all aspects of a spatial sound field. Then, these recorded signals are transferred to a remote location and are reproduced by means of spatial encoding and decoding, i.e. using ambisonic technique, through both headphones and loudspeaker arrays.

Alternatively, the sound sources are captured with close-up microphones alongside with metadata, such as sound source position, which may require a motion-capturing (mocap) system. This allows the audio rendering to take virtual room acoustics of the coupled rooms (cf. *The Immersion*) into account. Playback is conducted again through both headphones and loudspeakers.

Supplementing smart services, such as acoustic echo and feedback cancellation [5], [6] or online-estimation of room acoustic parameters support the immersion and are also covered by the LIPS Project.

Because the audio latency between the two sites is crucial for the musical interplay of the musicians, latency is a key topic of the LIPS research project. Methods are developed to measure the latency in audio transmission between two locations and two musicians, respectively [7]. Furthermore, strategies are investigated to enable the musical interplay also at slightly higher latencies [8], [4].

All methods described above rely on a robust IP-based connection to stream high quality audio at minimum latency.

In addition, specific hard- and software will be developed to provide a latency optimized gateway between the local audio production network and the wide area network and vice versa. The gateway adds as little additional latency and jitter as possible to facilitate in a low-latency connection between the two rehearsal rooms. Current devices on the market do not meet the latency requirements needed to enable interconnected rehearsals. It is planned to provide 6 microphones per room with a data rate of approx. 2.1Mbit/s per microphone. Each rehearsal room is prepared, so that the user can create his own audio monitor signal mixed from the 12 original microphone streams, 6 microphones in his local rehearsal room and 6 in the connected room. How to mix the microphone signals with different latencies to provide a useful monitor signal for the user, will be evaluated within LIPS.

Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag

The LIPS Vision

The Video

An immersive bidirectional image path with low latency and as outlined before is aimed. This relates to the camera as well as the display at each side.

The software and firmware of the cameras are specifically designed for LIPS and provide native All-IP functionality for image transfer, synchronization and control. Each camera sends its image stream and at the same time receives the remote image stream over IP. A codec inside the cameras is used for image compression and decompression. Decompressed images are passed to a large format display via a synchronous serial digital interface (SDI) socket. Both cameras are synchronized via the Precision Time Protocol (PTP) that is exposed as a Smart Service synchronously at both sides of the LIPS setup.

A ProRes codec compliant to the Apple ProRes standard is used. ProRes is designed for professional media production at the highest quality. It provides Wide Color Gamut and High Dynamic Range (HDR). ProRes is an “intra-frame” codec that allows for low latency. The implementation in the camera is capable of real-time compression of up to 200 frames per second (fps) which results in latency for compression and decompression below 10 milliseconds. The nominal bitrate for LIPS is about 650 Mbit/s in each direction.

The cameras are using the typical protocol stack for media production: Ethernet, IP, UDP, RTP, a media wrapper and the compressed video payload. The media wrapper is MXF Live according to SMPTE ST 2049. MXF is a KLV (Key, Length, Value, SMPTE ST 336) based encoding scheme that reliably identifies all payload by a unique key and without any additional control paths (technically spoken “in-band signaling”). The structure of MXF is applicable for file storage and IP streaming without any conversion process between. Data consistency between both domains can be achieved.

LIPS aims a calibrated, linear image path from scene to image reproduction. With respect to displays, this requires a defined and constant relation between a digital input value and the corresponding output luminance. SMPTE ST 2084 defines a suited relation over a wide dynamic range and with an efficient perception-based quantization. This standard is claimed to be supported by several high-quality displays. After testing various kinds of displays a large format OLED display has been selected. It is able to provide this characteristic in a constant and adequate way.

The display offers a low-latency mode. Values ranging from 12 to 15 milliseconds have been measured dependent on the position of the test pattern on the screen. This mode restricts the amount of internal signal processing. Regardless of that, the spatial and temporal image quality in this mode is good and without artefacts when providing an input stream of 50 frames per second and a clean full HD resolution as foreseen in the LIPS project. In standard operation mode latency is above 100 milliseconds. This would have significant impact on the overall latency budget and must be avoided to match the performance goals given in the annex of this paper.

The Light

Immersive integration of separate locations is essentially supported by harmonized lighting at both locations. Natural available light shall be maintained as much as possible to preserve a natural environment. To achieve this available light, the spectral power distribution of light at each location, is measured and processed by a central lighting control. A common target value for both locations with respect to lighting color coordinates and lighting intensity is calculated. The target is achieved

by adding artificial lighting. This is calculated specific for each site. Minimal additive lighting at both locations is aimed.

This process is executed continuously and maintains equal lighting conditions at both locations also when the available natural base light is changing. LIPS makes use of LED spectral multichannel lighting devices. These allow for a smooth approximation to the calculated target value in a wide range. Color coordinates and brightness can be matched precisely.

The Synchronization

To enable joint production at spatially separated locations, synchronization between PMSE devices such as microphones, video cameras, (in-ear) monitors, etc. must be established.

Regarding the LIPS use cases three different network deployments are discussed where synchronization between devices is mandatory. The first and the second cases are limited to one area, where all devices are locally available. In the first case, the data connection between different devices is established via cable, which can also be used for synchronization. In the second case, devices are connected wirelessly, e. g. via mobile communication like LTE-A or 5G New Radio. The third network deployment is the connection between spatially separated areas i.e. stages/rooms in different cities. The link in between is realized over a Wide-Area-Network. Synchronization between these areas has also to be established. All three cases offer different conditions for the achievement of synchronicity.

The requirement for the local synchronization is below 1 µs (cf. *The Annex*). For the wired case, all communication between devices is IP-based in the LIPS Use-Cases. Therefore, the implementation of the synchronization should also be IP-based to utilize the same physical cable for data and synchronization. In this case, IEEE 1588 Precision Time Protocol (PTP) [9] is used. To achieve best accuracy via PTP while network traffic is present, it is mandatory that the network between all synchronized devices support the PTP protocol and corrects time stamps accordingly. Considering wireless communication technology, the wireless transmission introduces time jitter and inconsistent transmission asymmetries between up- and downlink. Therefore, the IEEE 1588 protocol is not suitable to reach the requirement of 1 µs. Recently, 3GPP started to specify the transmission for 5G NR as a Time Sensitive Network (TSN) bridge [10]. Under this assumption, the synchronicity of below 1 µs may be achieved with the IEEE 1588 PTP, because the mobile transmission is seen as transparent bridge while asymmetries and jitter latencies are compensated within the 5G NR network.

The requirement for remote synchronization between different locations is below 1 ms (cf. *The Annex*). Regarding PTP, the achievable synchronicity strongly depends on the available network parameters between the two locations. Due to the distance between the locations, the network may be under the control of a third party, the network provider, and thus high time jitter and asymmetries might influence the transmission so that PTP is not suitable for the synchronization requirement. However, synchronicity of two separate locations can be stabilized using GPS.

The Future

The LIPS Project and its use-case push the boundaries of what is possible today within networked music performance. Technology convergence towards an all-IP based network including The Audio, The Video, The Light and The Synchronization is a key enabler to support a new level of Immersion.

Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag

The LIPS Vision

Technological requirements to support immersion on both sides of connected locations have been identified and the consortium works towards prototype implementation and verification of an all-IP based immersive production network.

5G promises to be a network of networks. 5G as an integrated technology will include vertical industries and its use-cases such as LIPS. It supports IP based end-to-end communication and offers new deployment possibilities (non-public networks). If 5G should meet all identified requirements, it could be one solution for interactive immersive services via one network.

The 3GPP Release 15, as the first 5G release, is not yet able to support the LIPS scenario completely, especially on the Radio Access Network (RAN). The URLLC use-case in 3GPP is not fully supported and the availability for equipment cannot be ensured. In the future, new 5G releases can cover the LIPS use-case in a variety of aspects, such as:

- the RAN could connect the production and end-user devices to a local serving cell
- the network slicing concept could ensure end-to-end quality for production services, with respect to latency, synchronicity, jitter and data rates
- a non-public-network in a nomadic fashion could be easily deployed and dismantled on an event basis
- a 5G backbone network could ensure the strict quality requirements between cells, location, rooms and devices in conjunction with the RAN
- the multi-access edge computing (MEC) could provide a standard interface for A/V related over the top services such as audio-mixing, video-re/encoding or live assisted listening

It is now the time to test and develop first solutions based on the offered possibilities.

The LIPS Vision

The Annex

The LAN

	System Parameter		Comment
Audio	Application latency	< 4 ms	Maximum allowable mouth-to-air latency at application level, includes interfacing and ADC / DAC
	Wireless transmission latency + transmission interval	< 1 ms	Latency that is introduced per link of the wireless communication system including the transmission interval of the audio data
	User data rate per audio link	150 kbit/s – 5 Mbit/s	Different user data rates per audio link need to be supported for different audio demands and different usage (mono or stereo)
	Reliability	99,9999%	The packet error ratio (PER) of the system shall be below 10-4 for a packet size corresponding to 1 ms audio data
	# of audio links	50 - 300	Simultaneous audio links: microphones and IEMs
	Service area	≤ 10.000 m^2	Event area, indoor and outdoor
	Synchronicity	≤ 1 µs	All wireless mobile devices of one local high quality network shall be synchronized at the application level within the specified accuracy
	User speed	≤ 300 km/h	
Video	Application latency	< 120 ms	An end-to-end delay from scene to display below this value is aimed. This includes at the sender side: sensor integration time, sensor readout, basic camera signal processing, image compression, IP packetizing, buffering for traffic shaping and error correction. This includes at the receiver side: IP unpacking, image decompression, 3-dimensional image processing for color and luminance, image output by SDI/HDMI, display latency.
	User data rate	650 Mbit/s average	Average value. Peak value up to 850 Mbit/s
	Reliability	99,999999%	The residual packet error ratio (PER) of the system on application level shall be below 3*10-9 for a packet size of 1280 bytes..
	# of video links	1	One bidirectional video link
	Synchronicity	1us	Target value for professional production in a local network (SMPTE).
Lighting	Application latency	< 50 ms	Target value for interactive manual lighting control
	User data rate	1 Mbit/s	Symmetrical data rate to and from all devices.
	Reliability	99,99999%	The residual packet error ratio (PER) per lighting device shall be below 4*10-8 for a packet size of 200 bytes.
	# of lighting links	10 for LIPS	Significantly more for real productions
	Synchronicity	1 ms	

Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag

The LIPS Vision

The WAN

	System Parameter		Comment
Audio	Application latency	< 25 ms	Maximum allowable latency at application level between two networks, includes interfacing and ADC / DAC.
	User data rate	10 Mbit/s	Inter-network communication with approx. 6 audio links per local network as foreseen in LIPS.
	Reliability	99.99999	
	Synchronicity	≤ 1 µs	All wireless mobile devices of both local high-quality networks shall be synchronized at the application level within the specified accuracy.
Video	Application latency	< 150 ms	= Value for LAN plus WAN Round Trip Time for error correction
	User data rate	650 Mbit/s average	Same as for LAN. Peak value up to 850 Mbit/s
	Reliability	99,999999%	Same as for LAN.
	# of video links	1	One bidirectional video link
	Synchronicity	1 ms	
Lighting	Application latency	< 150 ms	Max. latency for interactive manual lighting control
	User data rate	1 Mbit/s	Symmetrical data rate to and from all devices.
	Reliability	99, 99999%	Same as for Local area network..
	# of lighting links	10 for LIPS	Significantly more for real productions
	Synchronicity	1 ms	

The LIPS Vision

The References

[1] C. Rottondi, C. Chafe, C. Allocchio und A. Sarti, „An Overview on Networked Music Performance Technologies," *IEEE Access,* Bd. 4, pp. 8823-8843, 2016.

[2] ITU-T, „Vocabulary for performance and quality of service - Amendment 5: New definitions for inclusion in Recommendation ITU-T P.10/G.100," ITU-T, 07/2016.

[3] P. Le Callet, S. Möller und A. Perkis, „Qualinet White Paper on Definitions of Quality of Experience (2012)," European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, 2013.

[4] R. Hupke, L. Beyer, M. Nophut, S. Preihs und J. Peissig, „Effect of a Global Metronome on Ensemble Accuracy in Networked Music Performance," in *Audio Engineering Society Convention 147*, 2019.

[5] M. Nophut, R. Hupke, S. Preihs und J. Peissig, „Verfahren zur Multikanal Echokompensation in immersiv verknüpften Räumen," in *Fortschritte der Akustik DAGA 2019*, Rostock, 2109.

[6] M. Nophut, R. Hupke, S. Preihs und J. Peissig, „Multichannel Acoustic Echo Cancellation for Ambisonics-based Immersive Distributed Performances," in *Audio Engineering Society Convention 148*, 2020.

[7] R. Hupke, S. Sridhar, A. Genovese, M. Nophut, S. Preihs, T. Beyer, A. Roginska und J. Peissig, „A Latency Measurement Method for Networked Music Performances," in *Audio Engineering Society Convention 147*, 2019.

[8] R. Hupke, L. Beyer, M. Nophut, S. Preihs und J. Peissig, „A Rhythmic Synchronization Service for Music Performances over Distributed Networks," in *Fortschritte der Akustik DAGA 2019*, 2019.

[9] „IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems," *IEEE Std 1588-2008 (Revision of IEEE Std 1588-2002),* pp. 1-300, 2008.

[10] „3GPP TS 23.501 v16.1.0: 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2 (Release 16)," June 2019.