

Upmix and Format Conversion of Multichannel Audio – an Opportunity for AI-Based Breakthroughs?



Jean-Marc Jot
Virtuel Works
Principal



Gordon Wichern
MERL
Principal Scientist



Sunil Bharitkar
Samsung
Research Director



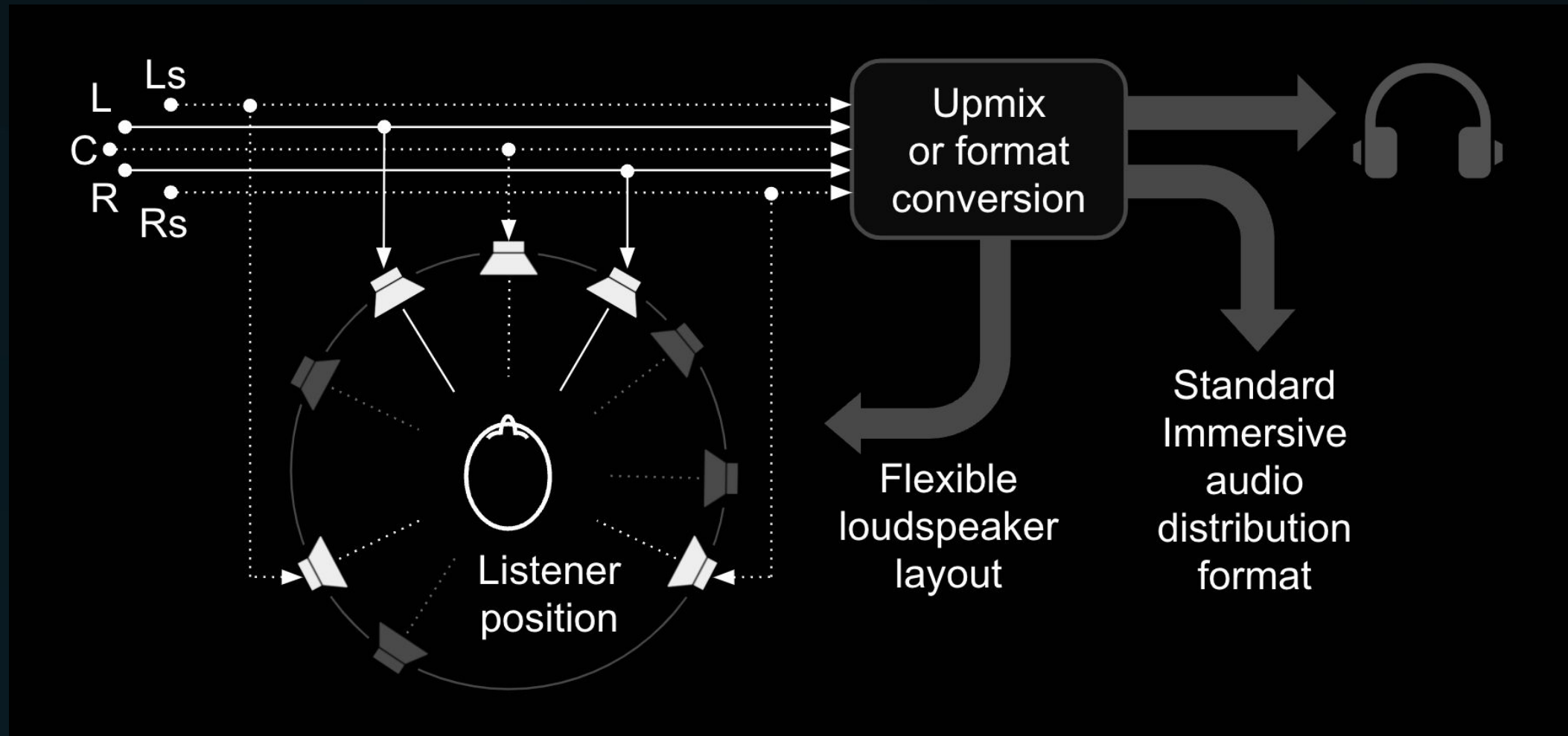
Carlos Freitas
Spatial9
CMO

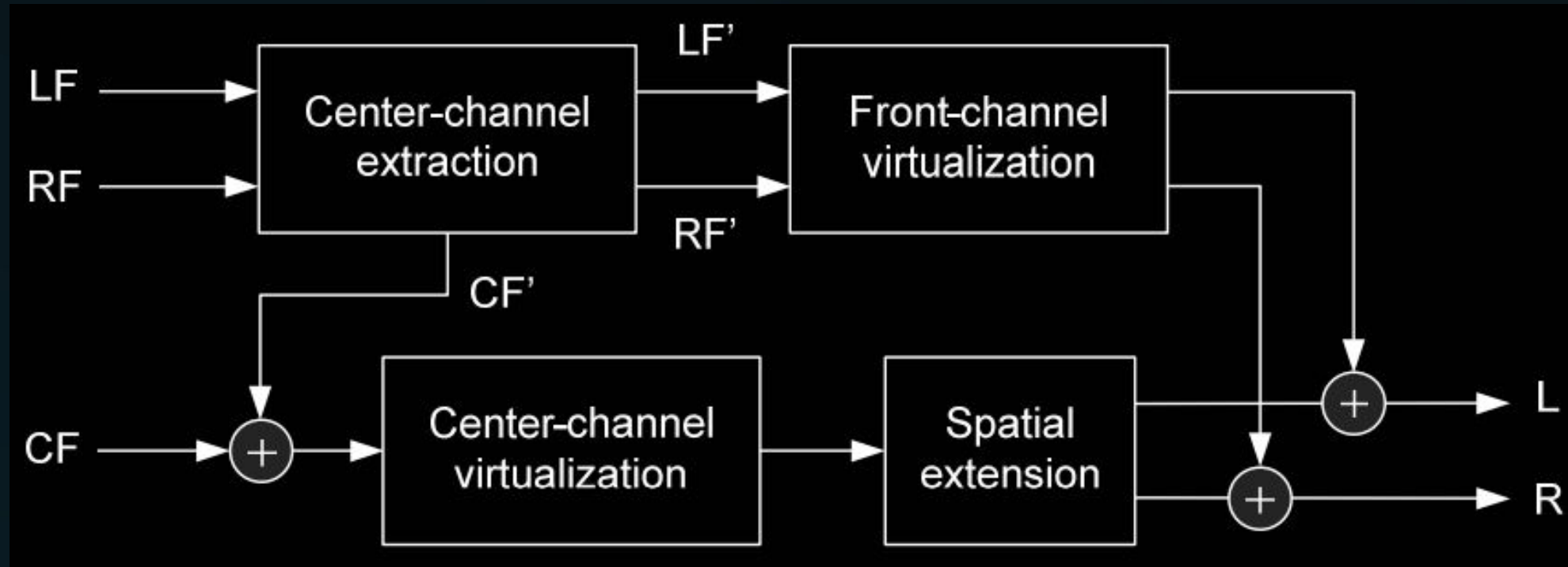


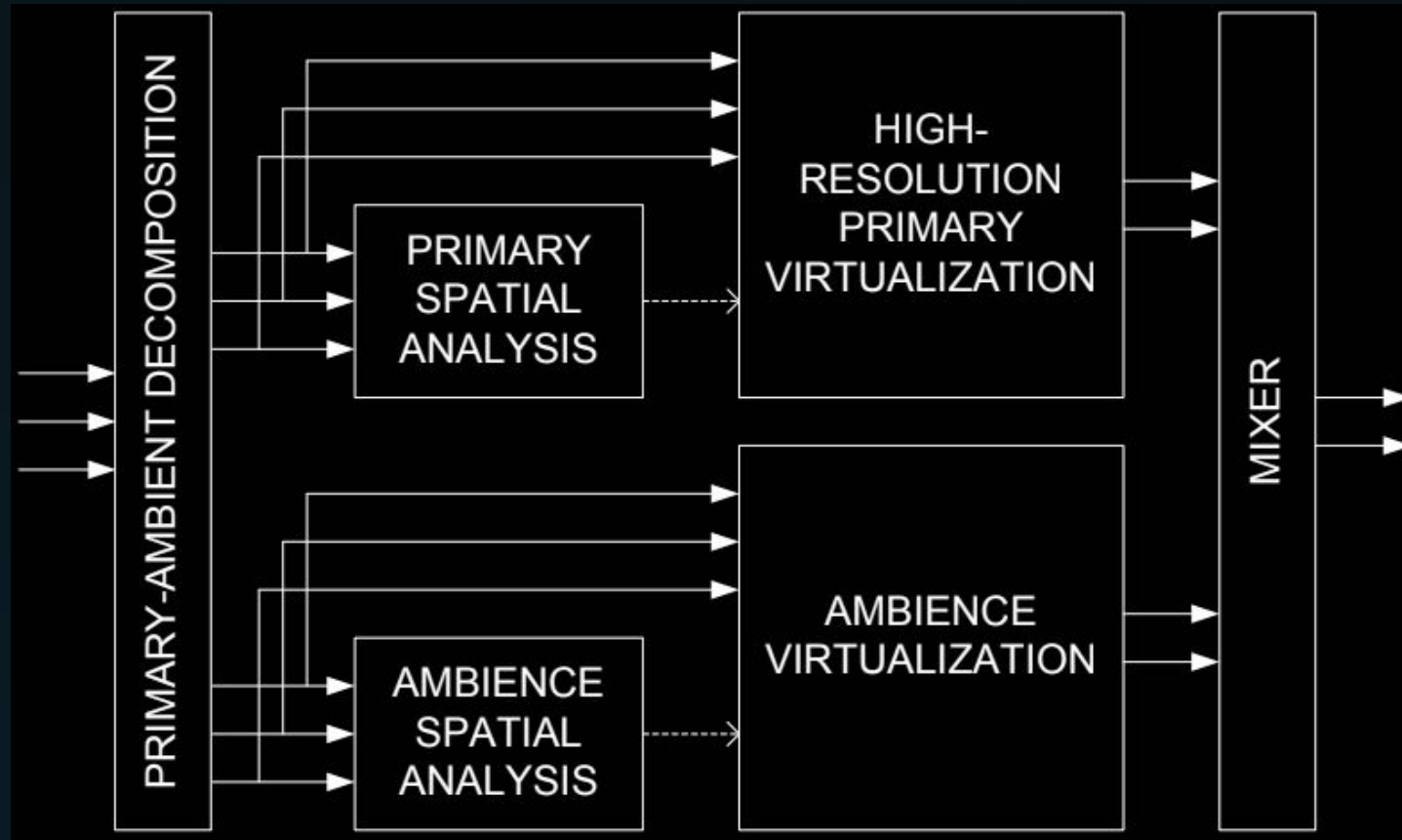
Alan Silva
Spatial9
CTO

Upmix and Format Conversion of Multichannel Audio – an Opportunity for AI-Based Breakthroughs?

Abstract: The demand for the production, distribution and playback of audio content in surround or immersive multichannel formats and flexible playback setups has continued to grow in the cinema, broadcast and music industries. This trend motivates the development of multichannel audio signal processing methods for the conversion of recordings between different multichannel and spatial audio formats or layouts, including down- or up-mixing. These operations leverage a well researched frequency-domain processing framework nowadays commonly referred to as parametric spatial audio signal processing, to solve challenges including the primary-ambient decomposition of audio recordings. Bringing ML/AI techniques to bear in the development of novel spatial audio analysis/synthesis methods can unlock practical applications with transformative impact. In this workshop, we will review and discuss the fundamentals and recent progress in these methods.







Machine Learning

Field of study that gives computers the ability to learn without being explicitly programmed.

(Arthur Samuel, 1959)



Machine Learning

Why/when is machine learning relevant/useful in audio signal processing?

- **DSP & math problems for which we don't have a practical analytic solution** (or are not confident we have a valid one)
 - offline (training): expose a computer to a large dataset of representative examples
 - runtime (inference): produce an estimate upon receiving a new input sample.
- **Ecological rationale / brain plasticity:** the human brain evolves / learns to take advantage of consistency among the stimuli that individuals are exposed to over many generations and over one's life experience (two different time scales).

Examples of specific tasks it can help us with?

- Stem separation/extraction from a recording
- Sound or effect matching, DSP algorithm parameter settings/preset search or recommendation
- Semantic understanding of a recording
- Production/mixing/mastering assistant.

Can Advances in Audio Source Separation Enable Better Upmixing?

Gordon Wichern

AES Show - NYC

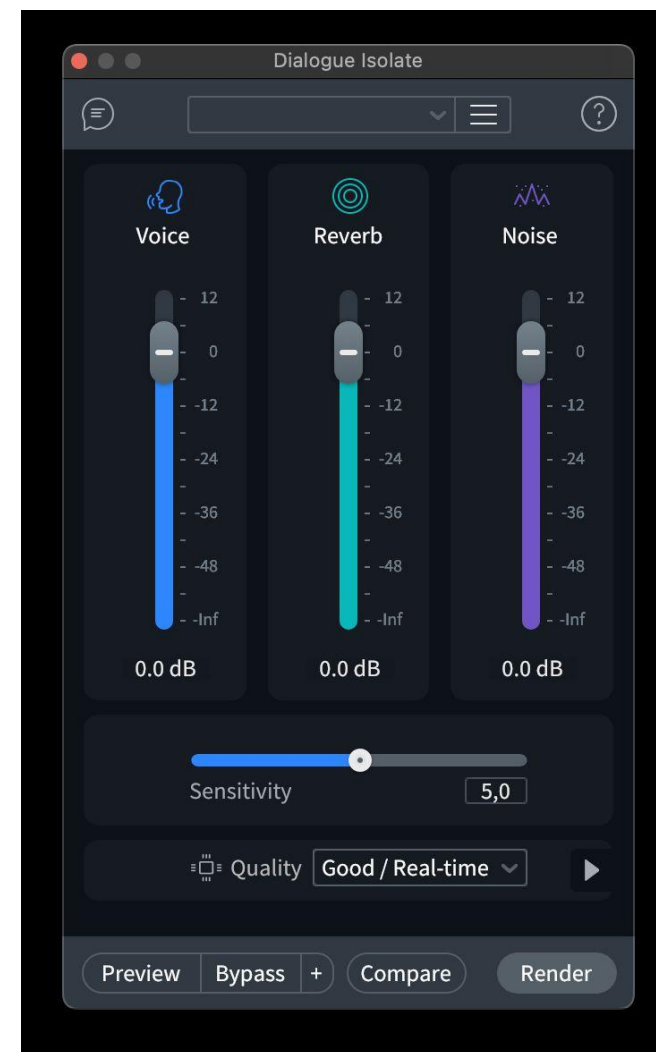
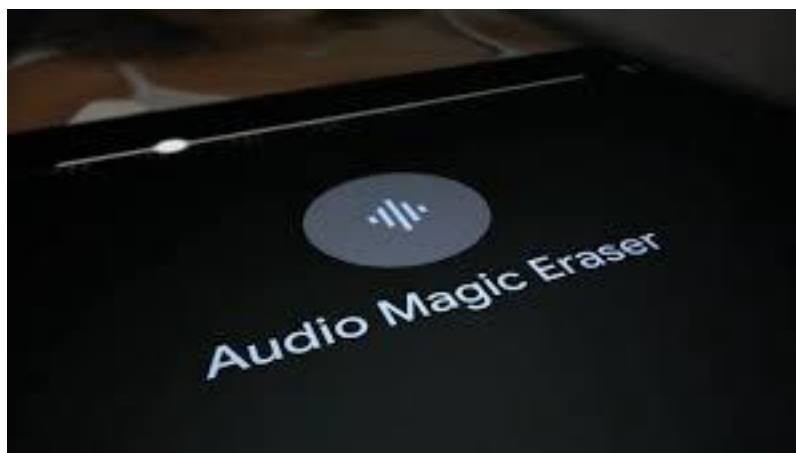
October 8, 2024

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)

Cambridge, Massachusetts, USA

<http://www.merl.com>

Audio Source Separation



Cinematic Sound Demixing

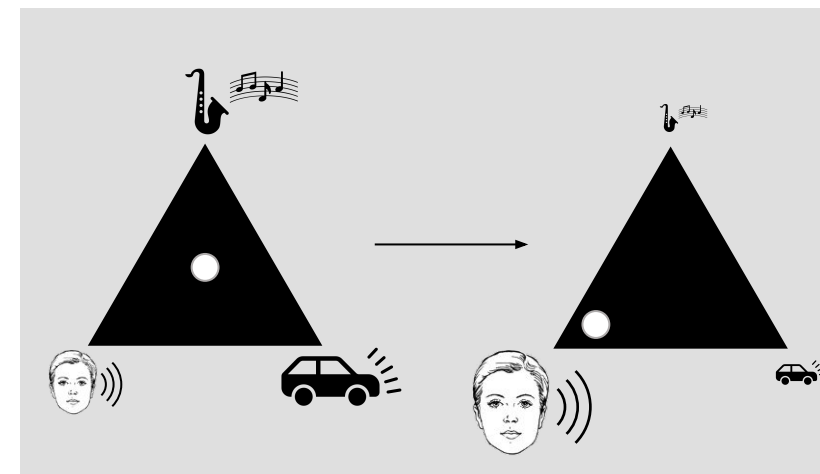
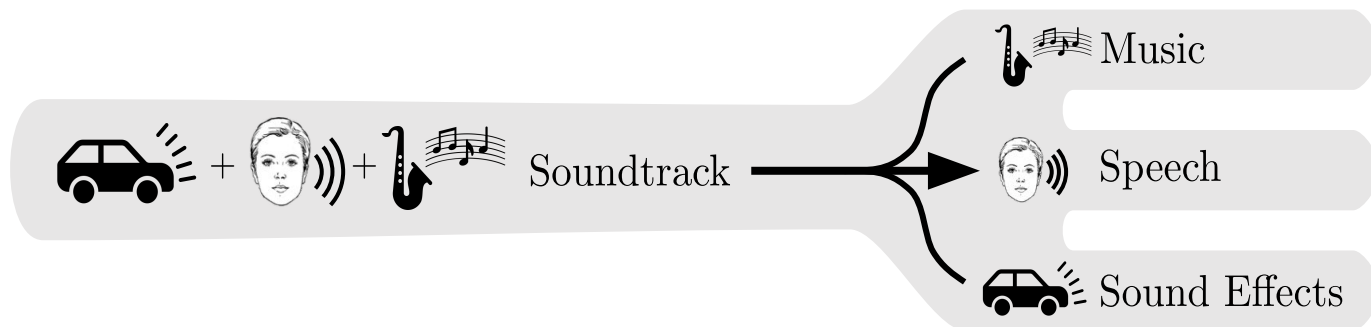
Sound Demixing Challenge 2023

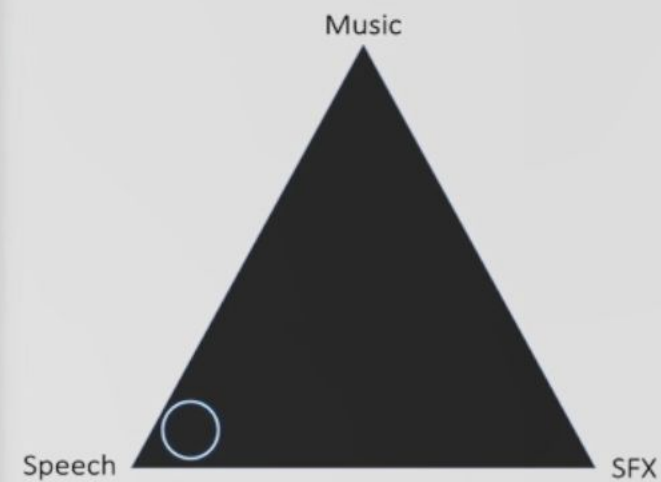
Build A System To Perform Audio Source Separation

 **\$42,000**
Cash Prize Pool

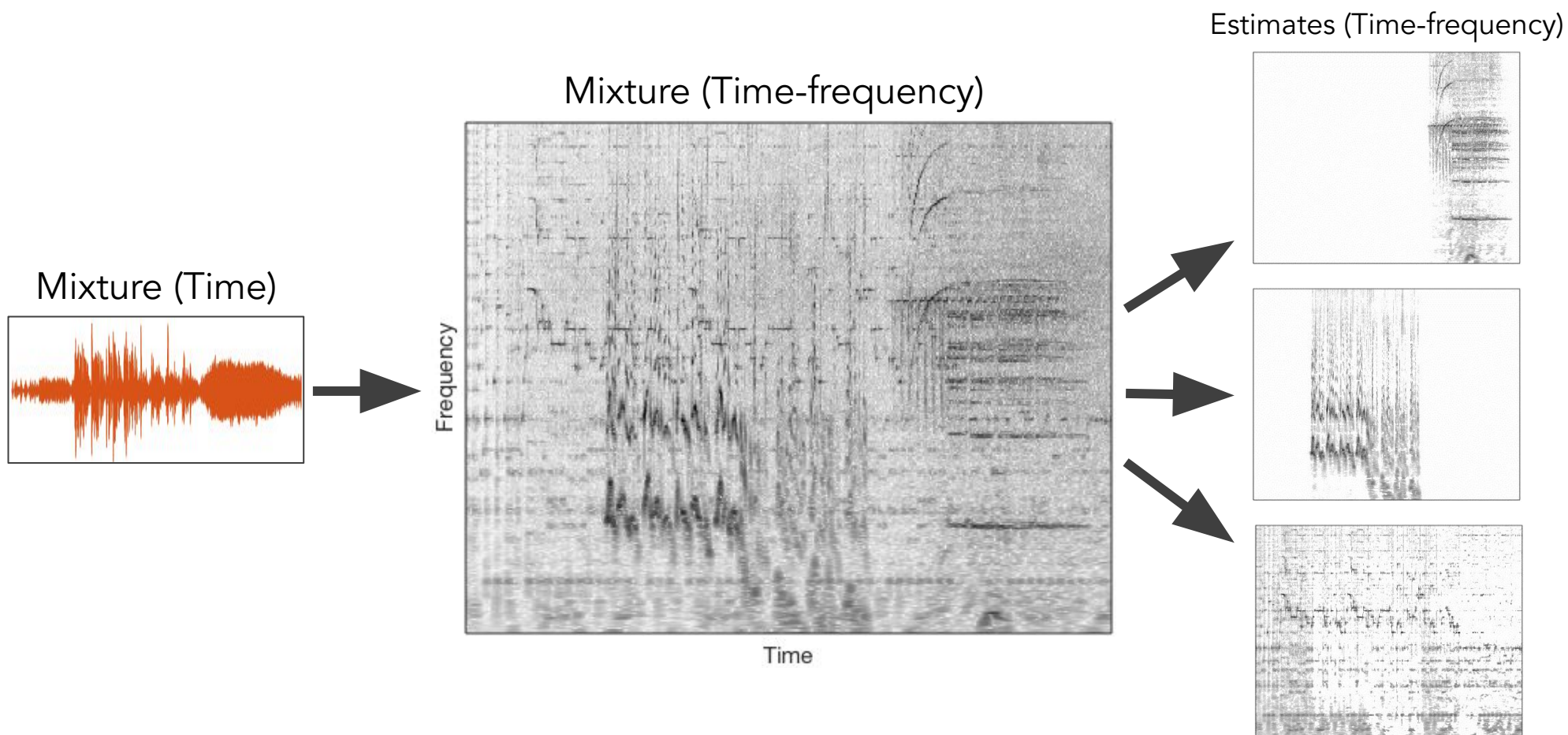


<https://www.aicrowd.com/challenges/sound-demixing-challenge>

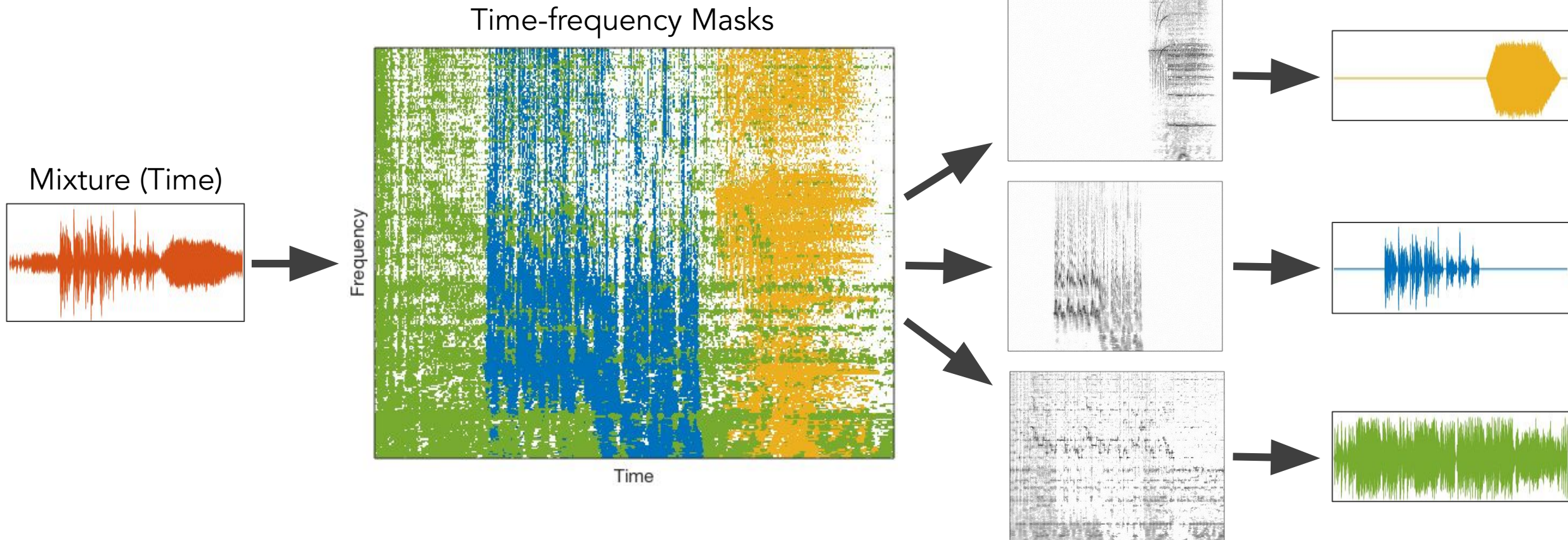
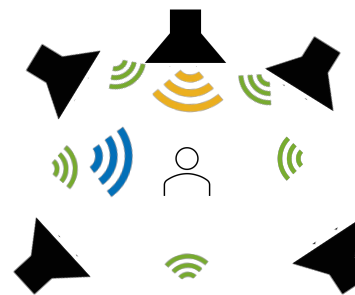




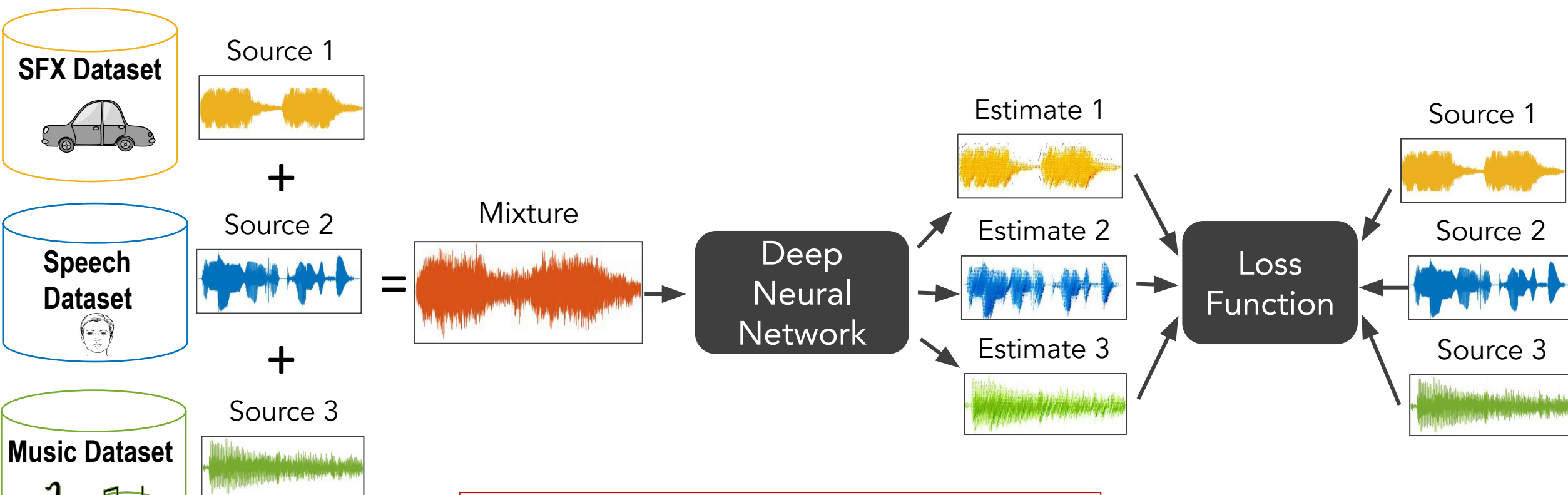
Time-Frequency Masks



Time-Frequency Masks

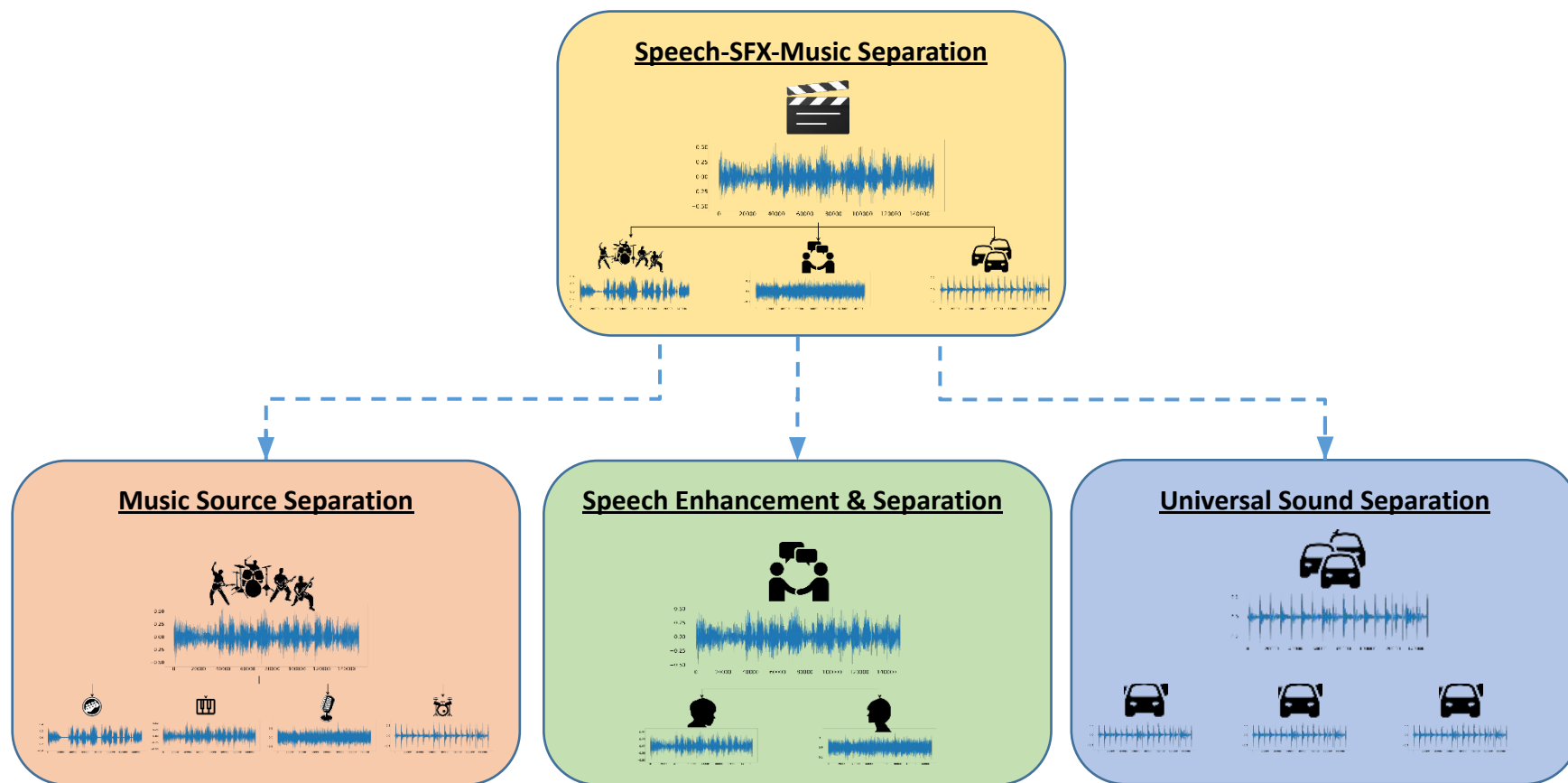


Deep Learning-based Source Separation Pipeline



- **Two Key Design Decisions:**
 - **Deep Neural Network**
 - Operate on waveforms
 - Operate on spectrograms
 - **Loss Function**
 - Time domain (waveform)
 - Frequency domain (spectrogram)

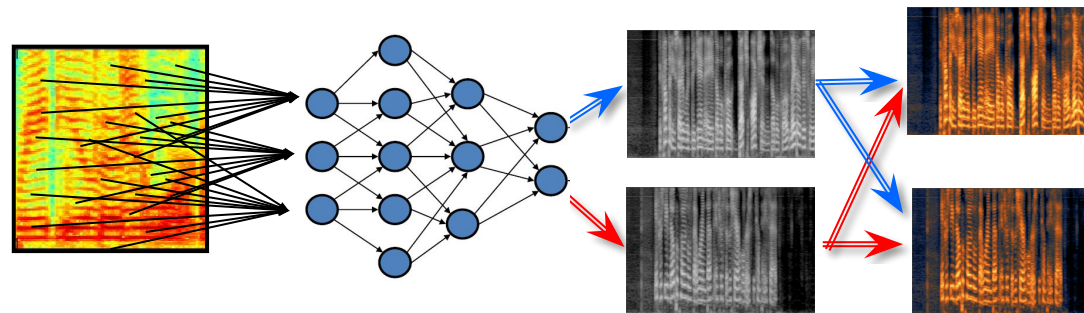
Source Separation Tasks



Important Advances for Flexible Source Separation

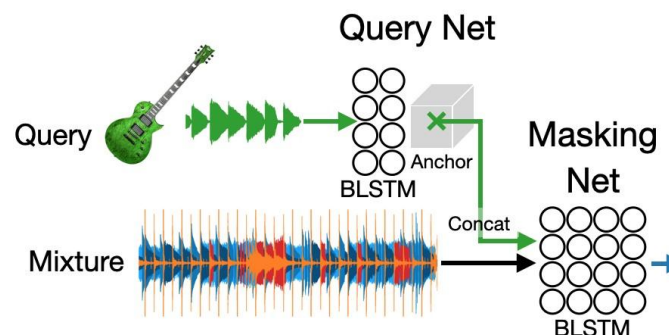
- Permutation-invariant training for separating multiple sources of the same type

- The speech from two speakers
- Any types of sounds



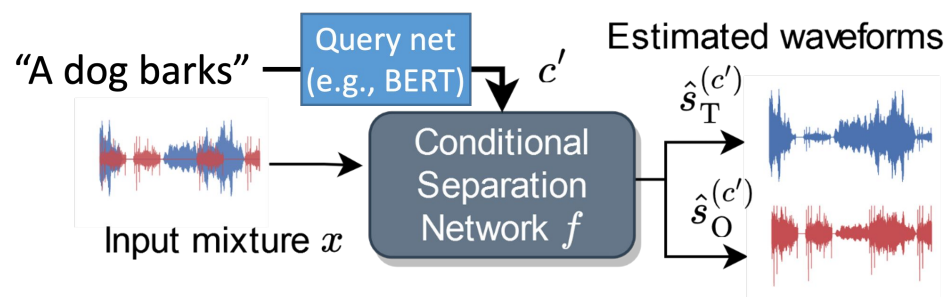
- Query-by-example based separation

- Go beyond fixed classes
- Target speaker extraction

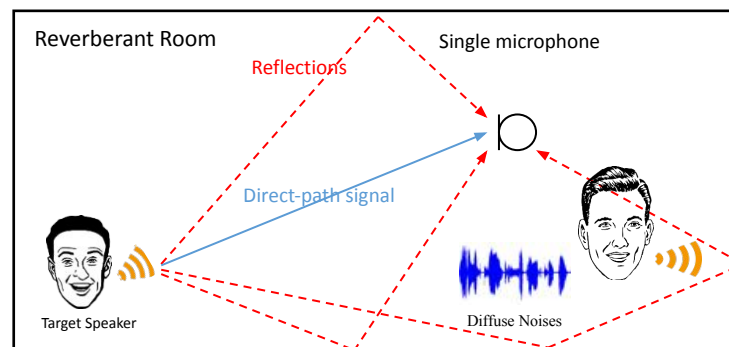


- Text-based separation

- Take advantage of advances in large language models

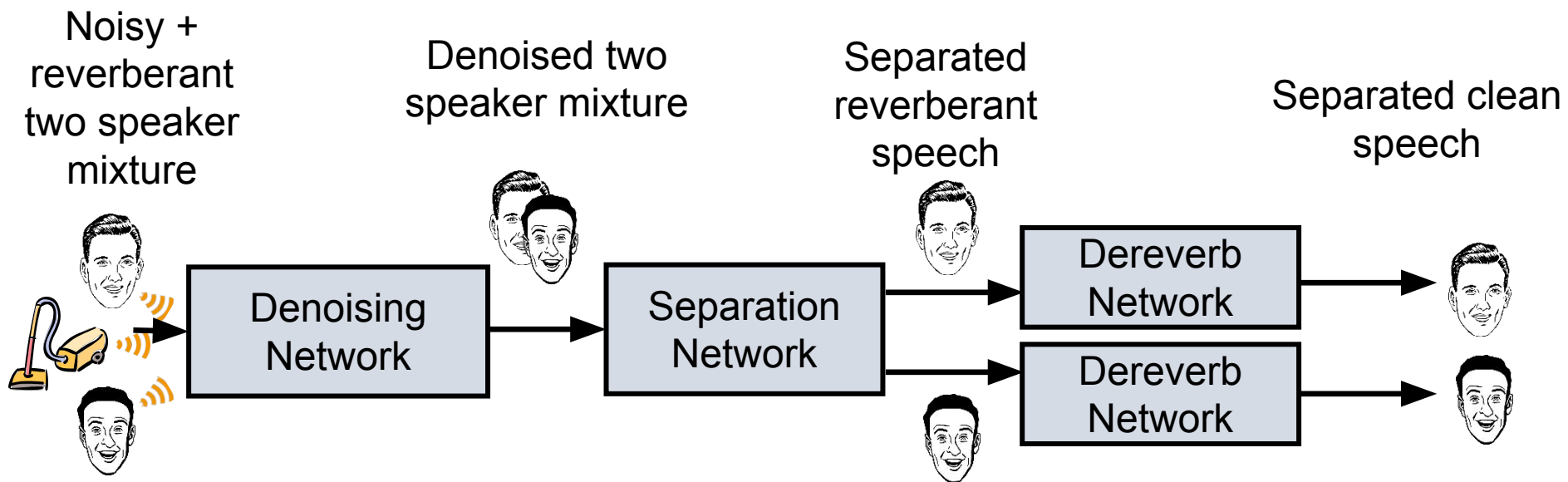


Separating Diffuse Noise and Reverberation



- Training requirements
 - A database of approximately clean anechoic sources
 - Reverberation simulator
 - Pyroomacoustics
 - gpuRIR
 - Noise dataset

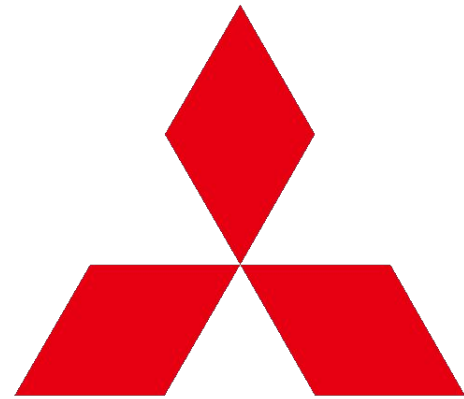
Separating Diffuse Noise and Reverberation – Cascade Approach



- First remove noise, then separate sources
- Because each speaker is located at a different point in space, they will have unique reverb, so we should apply dereverberation after separation

Recent Developments and Take-aways

- Don't separate, regenerate
- Unsupervised learning
 - Learn to separate sources with only mixtures
- Impulse response interpolation
 - Increasing attention in machine learning field
 - These approaches need input from people who focus on audio quality
- Takeaways
 - Source separation quality has increased dramatically with deep learning
 - The types of source we can separate and flexibility of tools is quickly increasing
 - Hopefully these advances can be used by next generation rendering engines to enable new upmixing research and tools



**MITSUBISHI
ELECTRIC**

Changes for the Better

AI Model Towards Immersive Upmixing

Sunil Bharitkar

Digital Media Solutions (DMS) Lab, Audio
Samsung Research America



Background

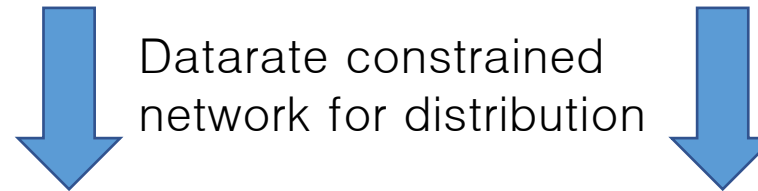
- Content Types
- Upmixing
- Preference

Content-dependent Classifier

- Classification
 - Taxonomy
- Classifier
 - Requirements
- Metadata
- AI Model
 - Results

Cinematic Immersive Upmixing

- Architecture
- Subjective
 - Testing
- Demo



- Content-dependency
- Processing-dependency

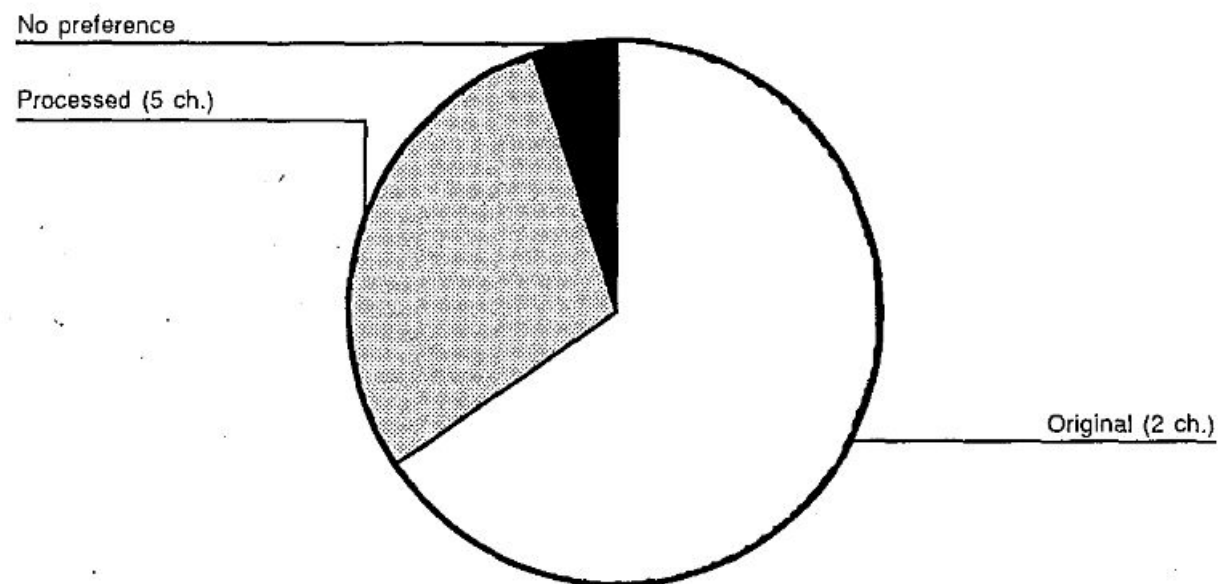


Fig. 10. Preference (all items).

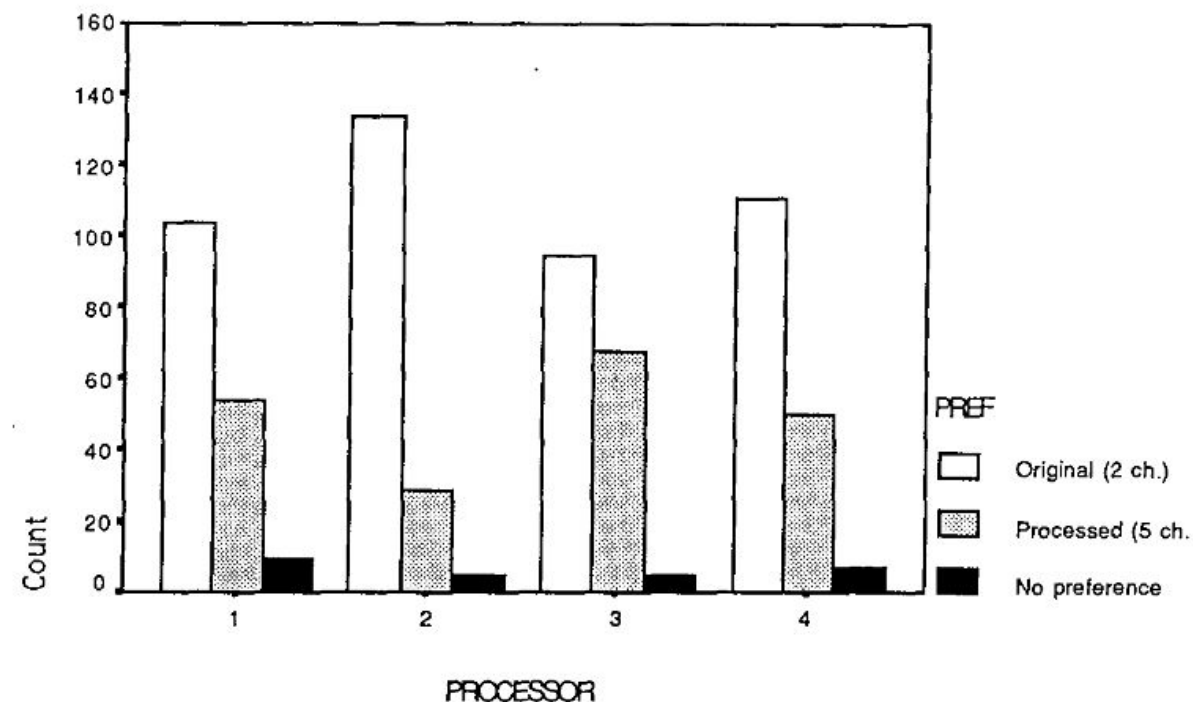


Fig. 12. Preference split by processor (all listeners).

- Content-dependency
- Processing-dependency

y

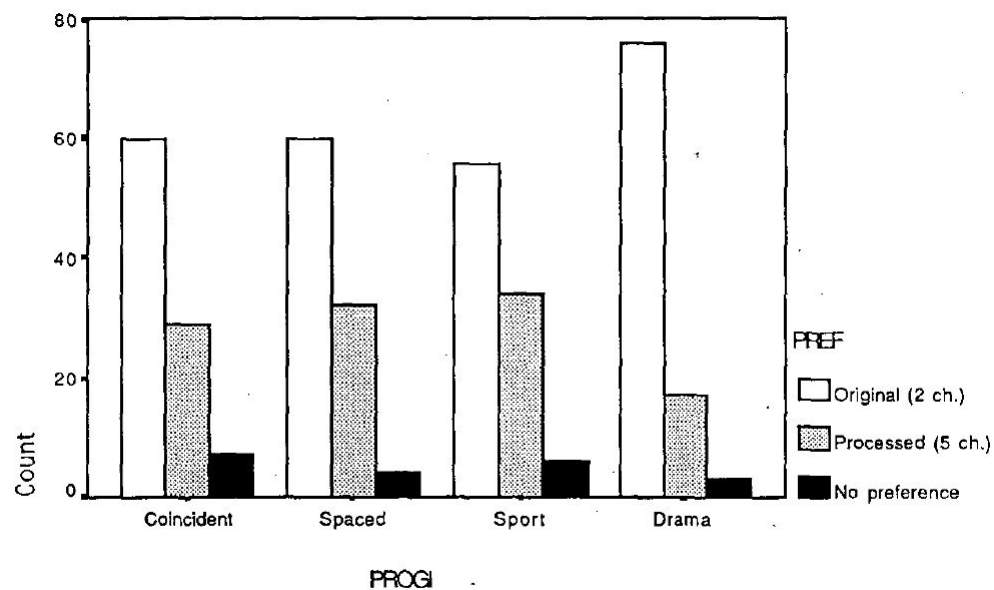


Fig. 15. Preference split by program type (all processors).

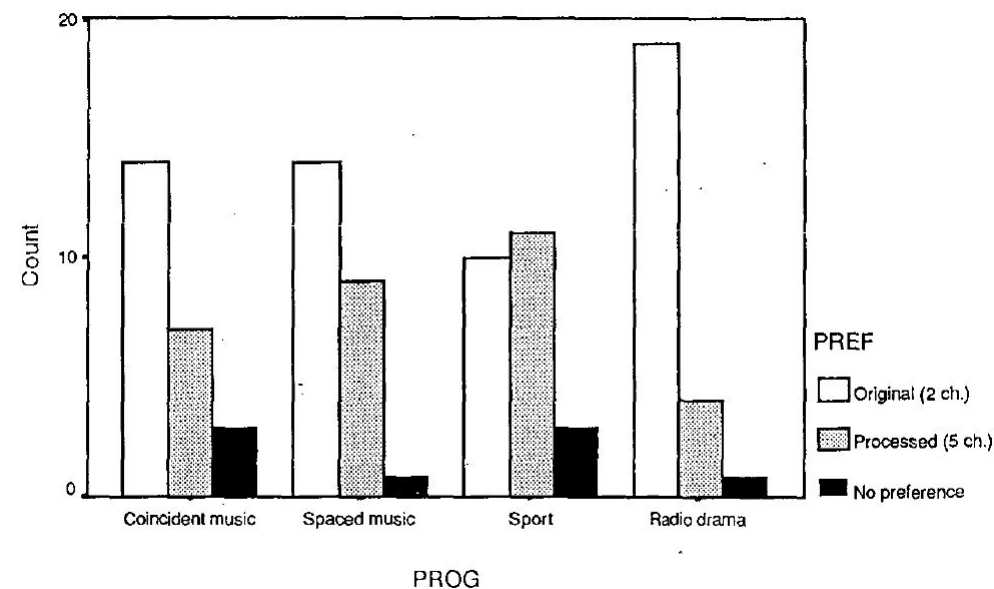


Fig. 16. Preference split by program type (processor 1).



Background

- Content Types
- Upmixing
- Preference

Content-dependent Classifier

- Classification
 - Taxonomy
- Classifier
 - Requirements
- Metadata
- AI Model
 - Results

Cinematic Immersive Upmixing

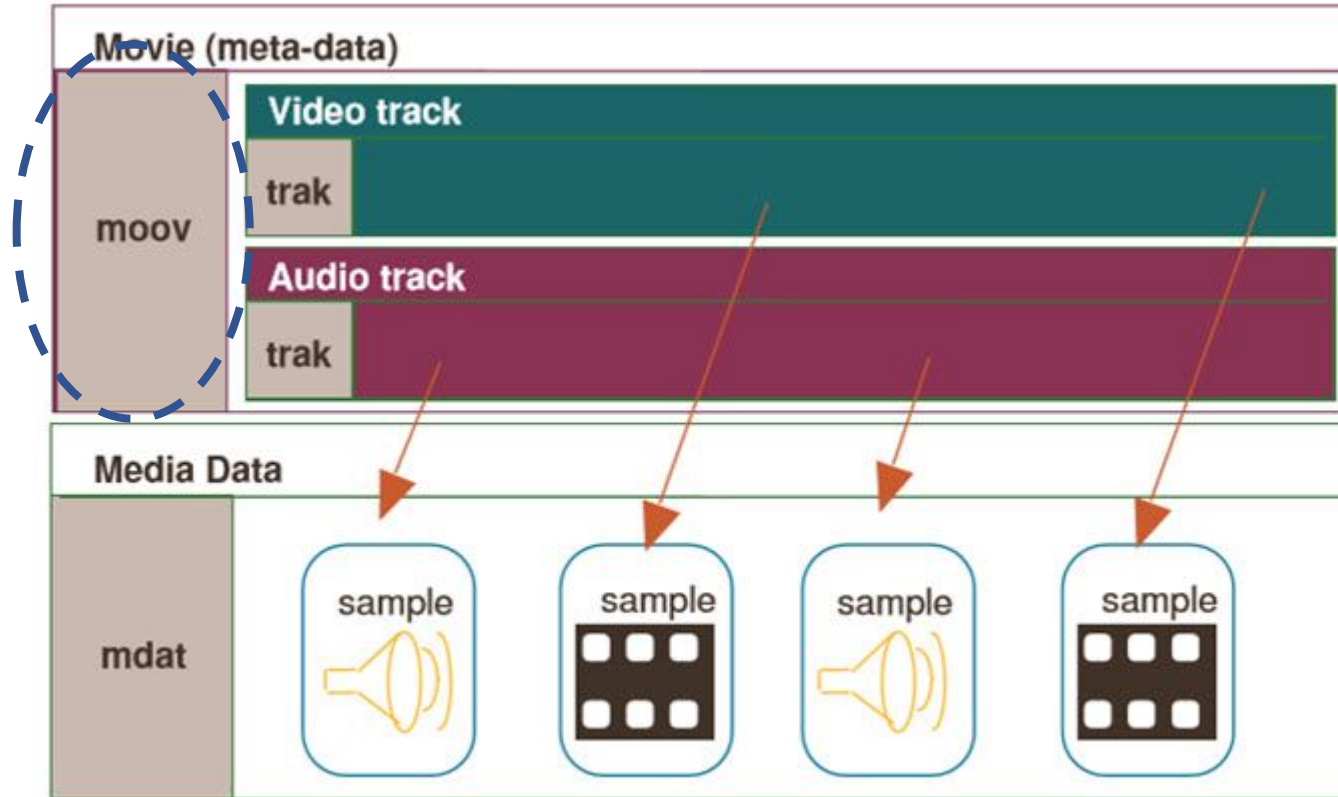
- Architecture
- Subjective
 - Testing
- Demo

- Classify mixed source streaming media content (e.g., YouTube) for audio or video post-processing on DSP/SoC
- Example (class-dependent audio post-processing):

Class	Processing
Documentary	PBE, Upmix, DE
Music	PBE
Movie	PBE, Upmix, DE
Live Music	PBE
News	DE
Sports	PBE, Upmix, DE
Education	DE
Ad	DE
Other	N/A

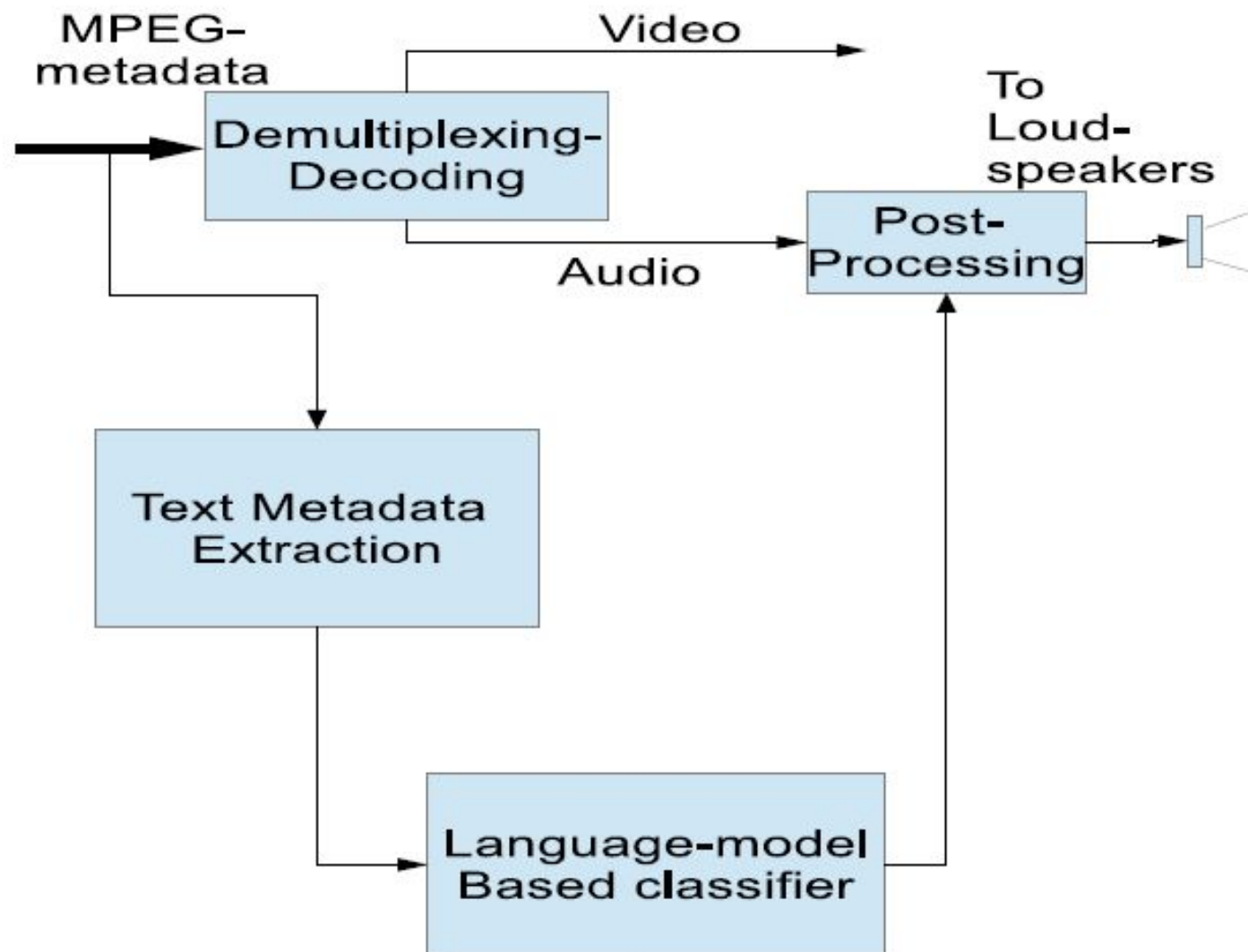
- Can we use non-audio or non-video data for classification on an edge device (smartphones, TVs, soundbars) satisfying:
 - fast response (decision)
 - low complexity
 - high accuracy
 - low-memory footprint
 - low/minimal battery consumption (mobile devices)
 - does not change classifier decision continuously (“fire and forget”) for audio or video enhancement

Can the *Title* metadata from the header of streaming media can be used for classification



Moov *header*: Holds all the movie/content metadata (including *Title*)

mdat: Contains media data

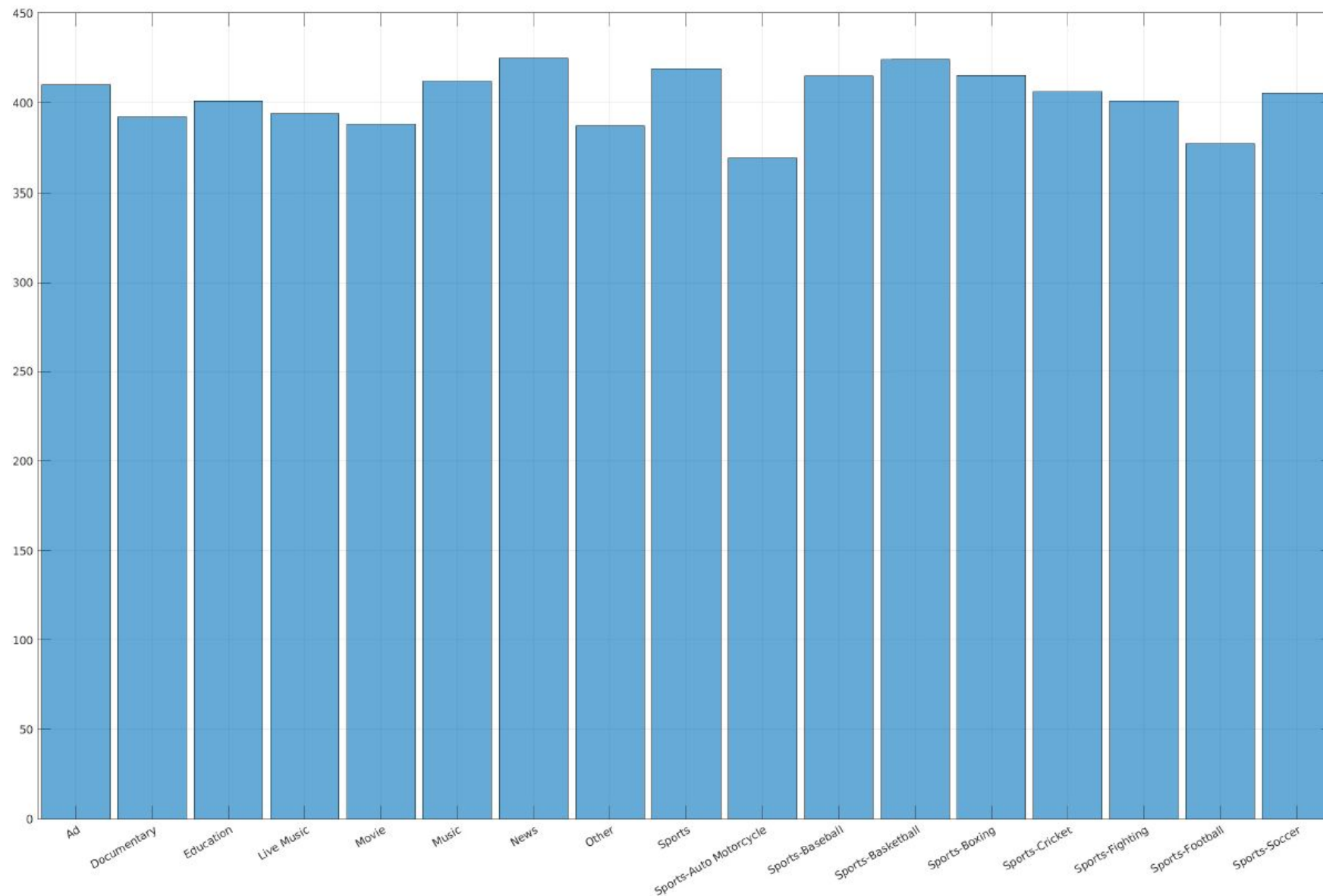


Example: 9 of 17

Class	Processing
Documentary	PBE, Upmix, DE
Music	PBE
Movie	PBE, Upmix, DE
Live Music	PBE
News	DE
Sports	PBE, Upmix, DE
Education	DE
Ad	DE
Other	N/A

YouTube | 8M

- 6840 YouTube *Title* metadata
- 17-class taxonomy



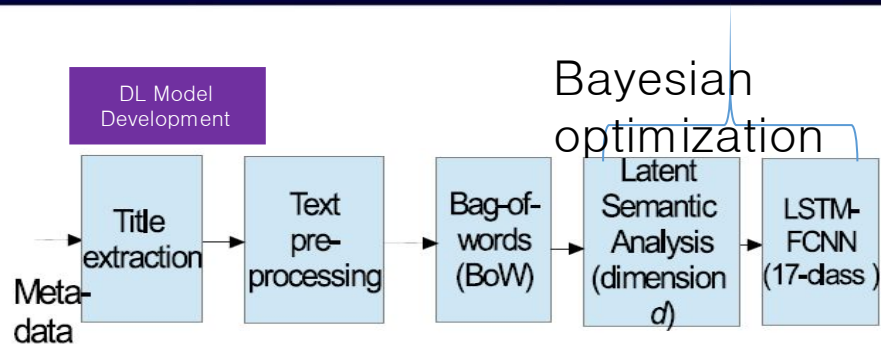


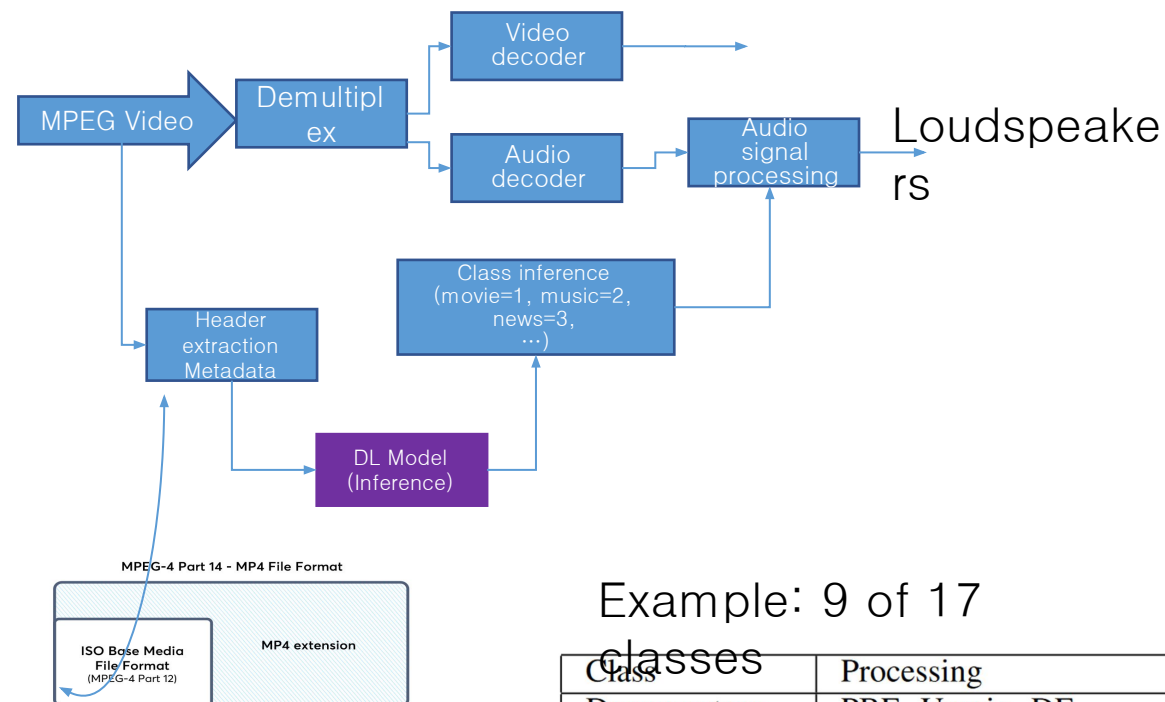
Fig. 2: Presented Bayesian optimized low-complexity and memory BoW-LSA-LSTM language model.

BoW

(example):

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

- 6840 YouTube *Title* metadata
- 75% training
- 5% validation during *bayesopt*
- 20% test



Example: 9 of 17

Class	Processing
Documentary	PBE, Upmix, DE
Music	PBE
Movie	PBE, Upmix, DE
Live Music	PBE
News	DE
Sports	PBE, Upmix, DE
Education	DE
Ad	DE
Other	N/A

Edge-optimized Model Algorithm

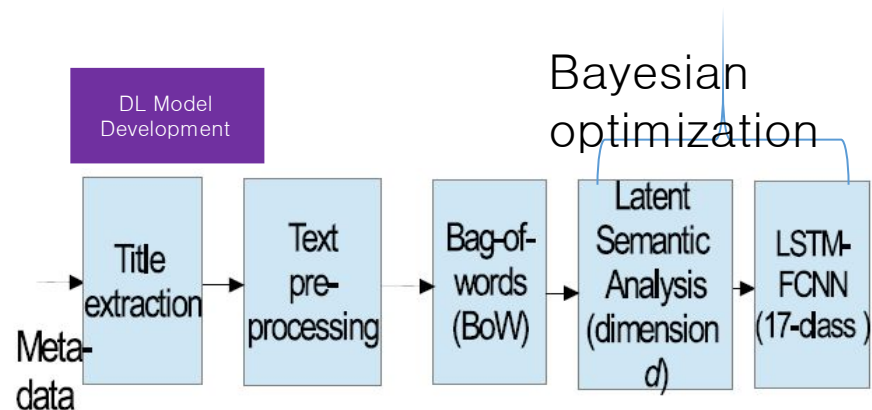


Fig. 2: Presented Bayesian optimized low-complexity and memory BoW-LSA-LSTM language model.

LSA: Low-rank factorization of BoW matrix using

$$\mathbb{R}^{5472 \times 10204} \stackrel{\text{SVD}}{=} \mathbb{R}^{5472 \times d_{latent}} \times \mathbb{R}^{d_{latent} \times d_{latent}} \times \mathbb{R}^{d_{latent} \times 10204}$$

LSA matrix used during inference

Algorithm 1: Bayesian Optimization (BO) for model search

Result: Best model for BoW-LSA(d_{latent}^*)-LSTM(M^*) having highest accuracy and F1 score over all 10-partitions of the test set $TS(p)$

- 1 Initialize: partition-count: $p = 1, 2, \dots, 10$, and Algorithm 2 early training-stop criteria (Max. of # Epochs = 800 and Duration = 5 hr. for each $p \in \{1, \dots, 10\}$);
- 2 while $p \leq 10$ do
- 3 Randomize training set $TR(p)$, validation set $VS(p)$, test set $TS(p)$;
- 4 Run Bayesian optimization (Algorithm 2) for $TR(p), VS(p)$;
- 5 Compute $\text{accuracy}_{\text{model}}(TS(p))$;
- 6 Compute $\text{F1score}_{\text{model}}(TS(p))$;
- 7 end
- 8 $\text{meanaccuracy}_{\text{model}}(TS) = \sum_{p=1}^{10} \text{accuracy}_{\text{model}}(TS(p))$;

$$M^* = 41; d_{latent}^* = 498$$

Algorithm 2: Bayesian hyper-parameter optimization

- 9 Initialize *bayesopt* search limits $d_{latent} \in \{5, 500\}$, $M \in \{10, 600\}$ (M defined in (1)), LSTM(M) parameters: ADAM optimizer, $\eta_0 = 0.1$, $\eta_{\text{drop-factor}} = 0.99$, $\eta_{\text{drop-period}} = 5$;
- 10 Create bag-of-words for $TR(p)$;
- 11 LSA(d_{latent}) using singular value decomposition on bag-of-words;
- 12 Train LSTM(M);
- 13 Generate $VS_{out}(p)$ on $VS(p)$ with model LSA(d_{latent})-LSTM(M); $VS_{out}(p) = \{\hat{e}_c\}$ is the set of unit vectors for the validation set $VS(p)$, each of dimension seventeen with zeros except at the c -th class position classified by LSTM(M) $c = \{1, \dots, 17\}$;
- 14 Use *bayesopt* to minimize $\text{Err}VS(p) = \frac{1}{|VS(p)|} \left(\sum_{r=1}^{|VS(p)|} 1[VS_{out}(r) - VS_{true}(r)] \right)$, where 1 is the indicator function, $|\cdot|$ is the cardinality of the validation set, $VS_{true}(r)$ is the true class for the r -th datapoint in the validation set;

TABLE I: Comparisons: Language Model Performance over seventeen-classes

Metrics	BERT _{base/tiny}	RWKV	RetNet _{mini/tiny}	Prop.
Test acc.*	88.4 / 78.9	84.8	97.59 / 96.49	86.1
F1 score*	88.4 / 78.9	84.8	97.59 / 96.49	86.3
Avg. acc.	86.2 / 77.1	79.77	94.7 / 91.9	84
Latency* (ms)	150 / 16.9	483.9	27 / 15	2.9
# param.* (Millions)	110 / 4.4	381.5	10 / 4.8	0.088



Sun *et al.*, “Retentive Network: A Successor to Transformer for Large Language Models,”
[arXiv:2307.08621](https://arxiv.org/abs/2307.08621), 2023

Fig. 4: Proposed BoW-LSA(498)-LSTM(41) implementation on a CPU in a TV



Background

- Content Types
- Upmixing
- Preference

Content-dependent Classifier

- Classification
 - Taxonomy
- Classifier
 - Requirements
- Metadata
- AI Model
 - Results

Cinematic Immersive Upmixing

- Architecture
- Subjective
 - Testing
- Demo

- Generalization of Equal-Levels Ambience Extraction [Merimaa, Goodwin & Jot, 2007]

- 1. Signal model $\vec{X} = \vec{P} + \vec{A}$
- 2. Equal levels $\|\vec{A}_L\| = \|\vec{A}_R\| = I_A$
- 3. Uncorrelated $\vec{P}$ and $\vec{A}$, as well as $\vec{A}_L$ and $\vec{A}_R$
- 4. Ambience soft-mask $\vec{A}(t, f) = \alpha(t, f)\vec{X}(t, f)$

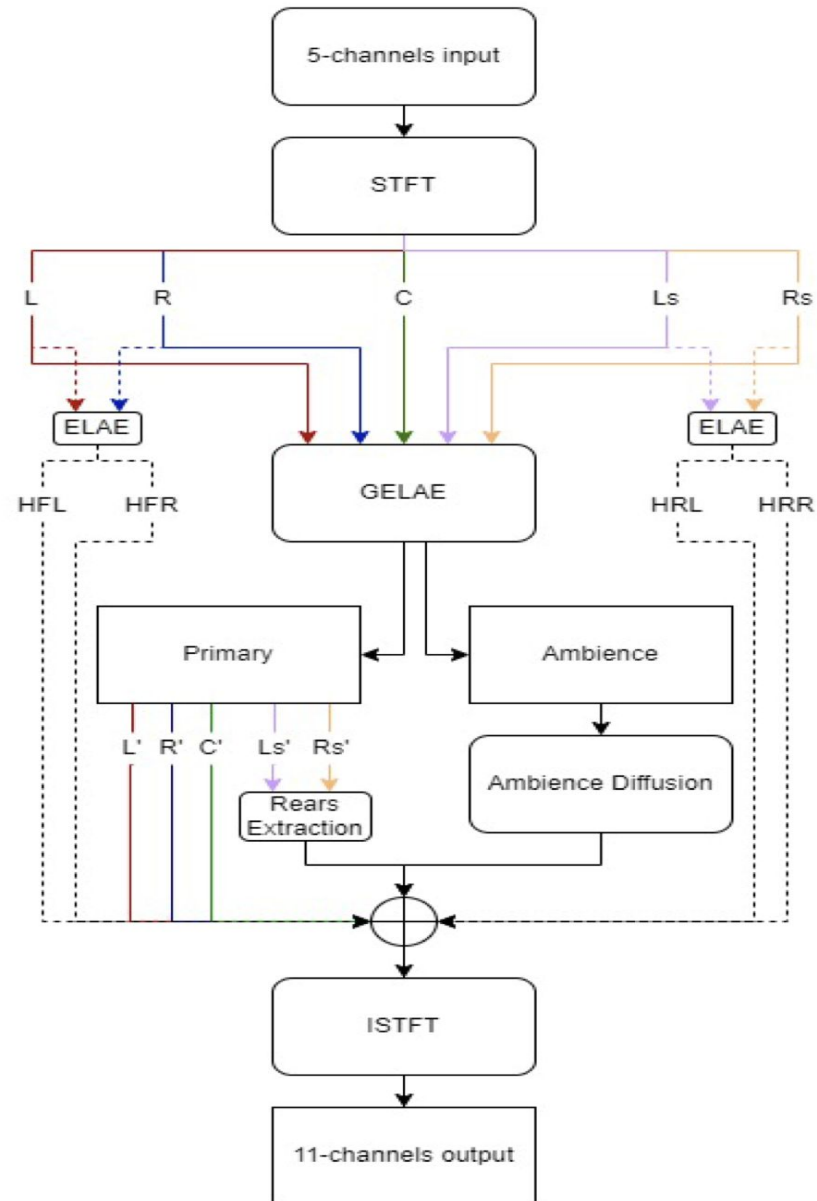
$$I_A^2 = \frac{1}{2} \left(r_{LL} + r_{RR} - \sqrt{(r_{LL} - r_{RR})^2 + 4|r_{LR}|^2} \right)$$

$$\alpha_L = \frac{I_A}{\|\vec{X}_L\| + \epsilon} = \frac{I_A}{\sqrt{r_{LL}} + \epsilon},$$

$$\alpha_R = \frac{I_A}{\|\vec{X}_R\| + \epsilon} = \frac{I_A}{\sqrt{r_{RR}} + \epsilon}$$

- Extending 2. $\|\vec{A}_i\| = I_A \quad i = 1, \dots, N.$

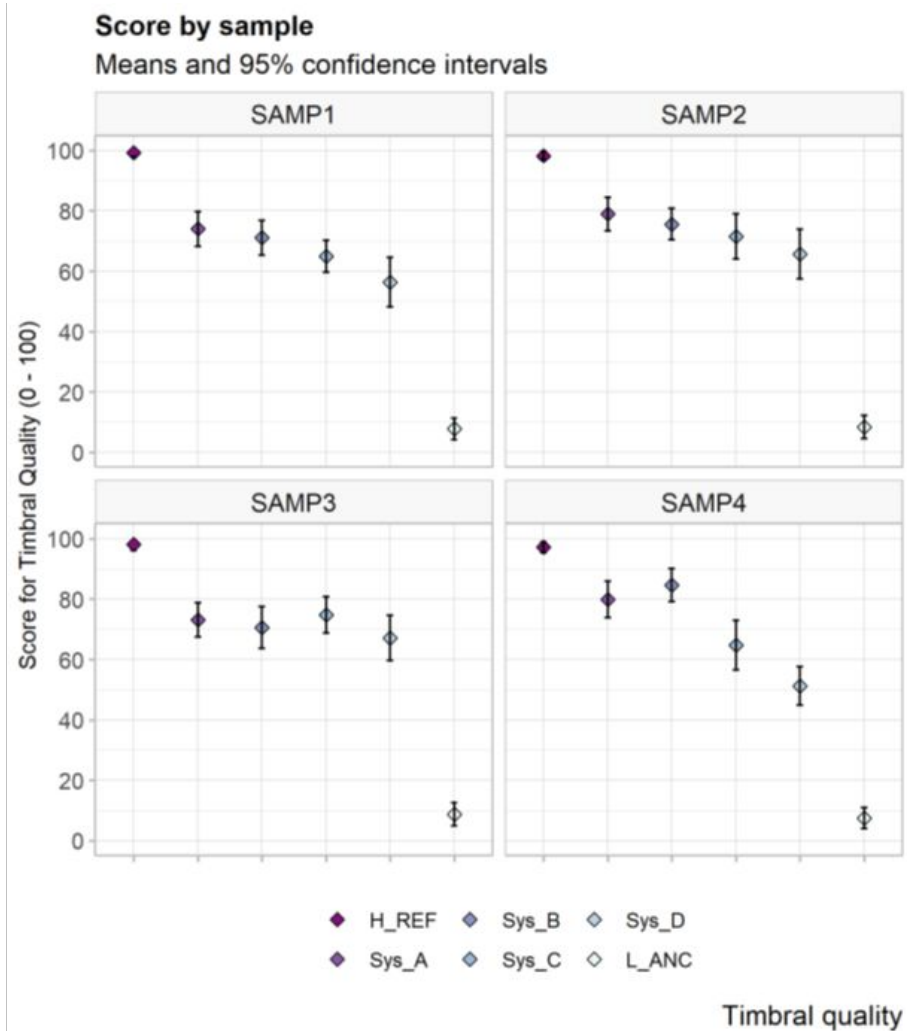
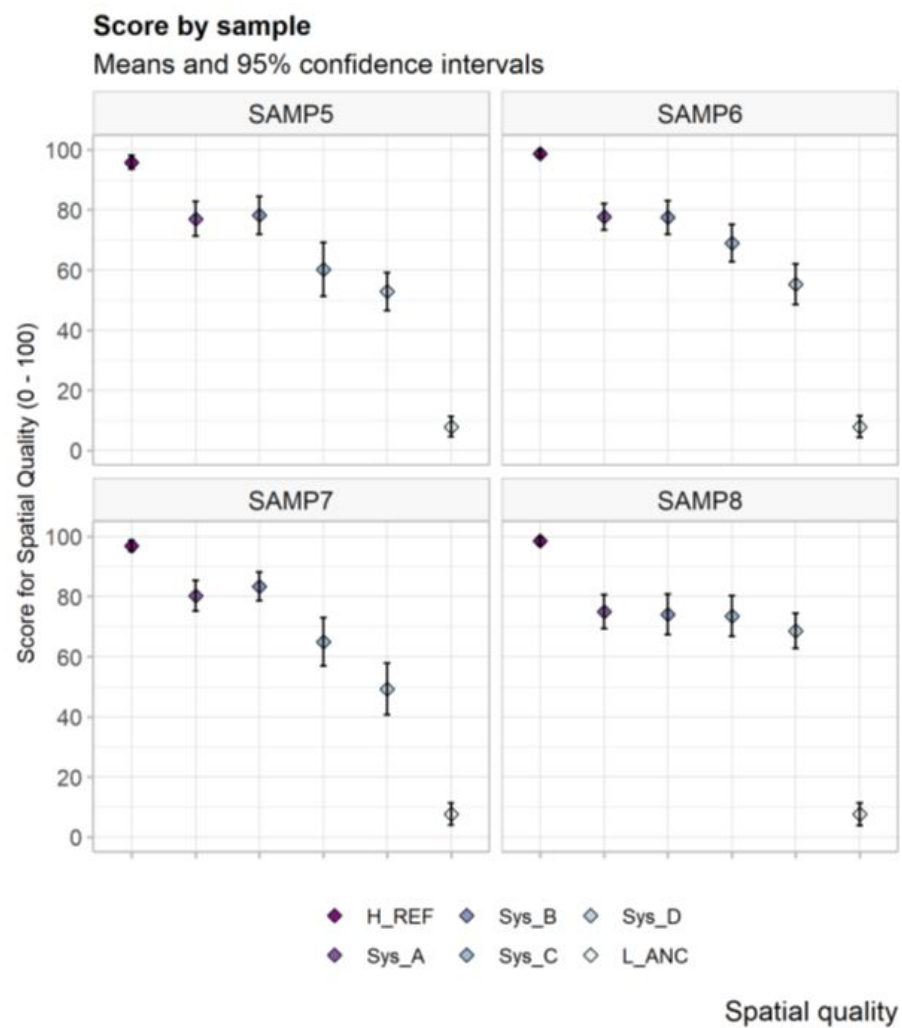
$$I_A^2 = \frac{1}{\binom{N}{2}} \left[\sum_{i=1}^{N-1} \sum_{j=1}^N \left(2r_{ii} - \frac{1}{2} \sqrt{(r_{ii} - r_{jj})^2 + 4|r_{ij}|^2} \right) \right]$$

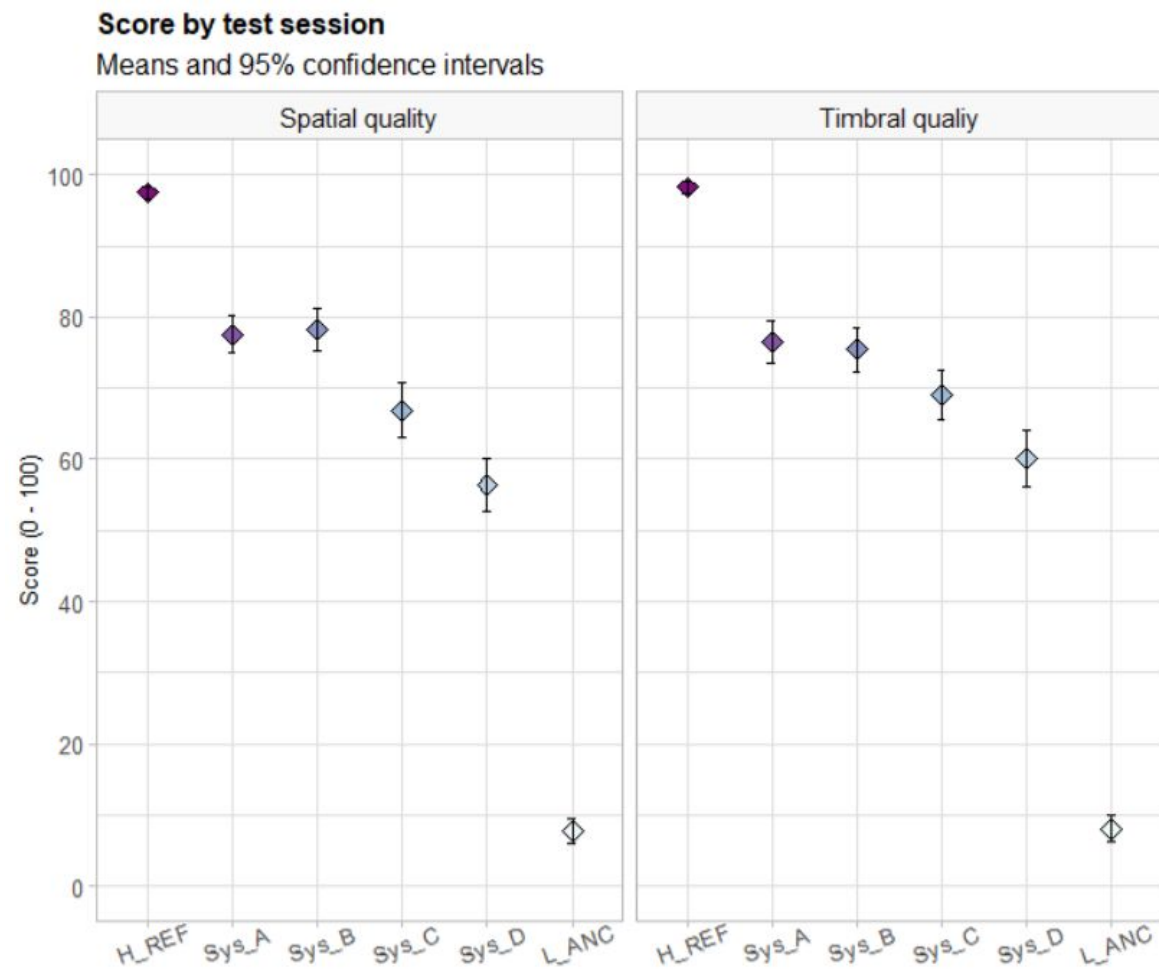


- Modified MUSHRA
 - Hidden reference
 - original 7.1.4 tracks
 - Proposed method
 - Variation (no stereo ambience)
 - Dolby Digital Surround upmixer
 - DTS Neural:X upmixer
 - Low anchor
 - 3.5kHz low-pass filtered version of reference, downmixed to stereo

Description of test tracks

		Label	Content	Dur.
Timbral	{	SAMP1	Concert scene. Male singer voice, with surrounding crowd clapping and singing.	10s
		SAMP2	Gathering of people cheering, with surround motion of firework.	32s
		SAMP3	Children playing in a forest, with sound effects, music and narration voice.	15s
		SAMP4	Steam train whistles and sound effects.	15s
		SAMP5	Family members arguing in a reverberant space, with steps and sound effects heard moving around and upstairs.	21s
	{	SAMP6	Children running around in a forest, with elevated narration voice.	15s
		SAMP7	Steam train whistling and getting in motion with many sound effects.	15s
		SAMP8	Army marching with horn blaring.	21s





Average results of objective tests

Upmixer	SI-SDR↑	MSE (10^{-4}) ↓	LSD (10^{-1}) ↓
Sys _A	-1.69 ± 1.68	2.27 ± 2.04	1.26 ± 0.14
Sys _B	-3.36 ± 2.34	2.28 ± 2.15	1.28 ± 0.16
Sys _C	-9.05 ± 2.27	3.78 ± 2.64	1.33 ± 0.09
Sys _D	-8.7 ± 2.3	3.96 ± 1.36	1.33 ± 0.08

- Upmix signals while retaining artistic intent
- Cinematic content can be upmixed with proper processing to give high preference scores
- Demo
- Acknowledgements: Thaddeus Paez (SRT) for joint-contributions to this work and Matthew McDuffy (SRA)

for

preparing demo material in ADM file format

Thank You



SPATIAL⁹

Simplify Spatial Sound

Alan Silva (alan@spatial9.ai)

Carlos Freitas (carlos@spatial9.ai)



The world of music is **changing**

The music industry is undergoing a groundbreaking transformation, fueled by the power of 3D immersive technologies and innovative audio formats driven by Big Tech companies and Music Streaming Services. Modern devices are unlocking new dimensions of sound, allowing artists to create immersive experiences that transport listeners into entirely new realms. This revolution is not just reshaping how we hear music—it's redefining how we feel and experience it, pushing the boundaries of creativity and innovation like never before. The future of music is here, and it's more dynamic, immersive, and awe-inspiring than ever.

From mono to stereo and now to immersive soundscapes, music has evolved into an emotional, vivid, and unforgettable journey.

The **PROBLEMS** of a changing world



HIGH PRODUCTION COSTS



LONG DELIVERY TIME



POOR RESULTS



TECHNICAL DEBT



HIGH DEMAND



COMPLEXITY

Solution Overview



IMMERSIVE AUDIO EXPERIENCES

Spatial9 develops cutting-edge audio technology that creates fully immersive, 3D sound environments for music.



EASY INTEGRATION

Spatial9's developer-friendly platform empowers artists and creators to mix and produce music in a true immersive format.



INNOVATIVE PARTNERSHIPS

The company's proprietary, advanced spatial audio algorithms deliver high-fidelity, multi-dimensional sound in partnership with renowned Studios and Google.



FAST AND AFFORDABLE

Spatial9's affordable technology empowers artists to produce captivating, immersive music in minutes with the cost of less than a few streams.

ARTISTS

PRODUCTION

DISTRIBUTION

CONSUMPTION

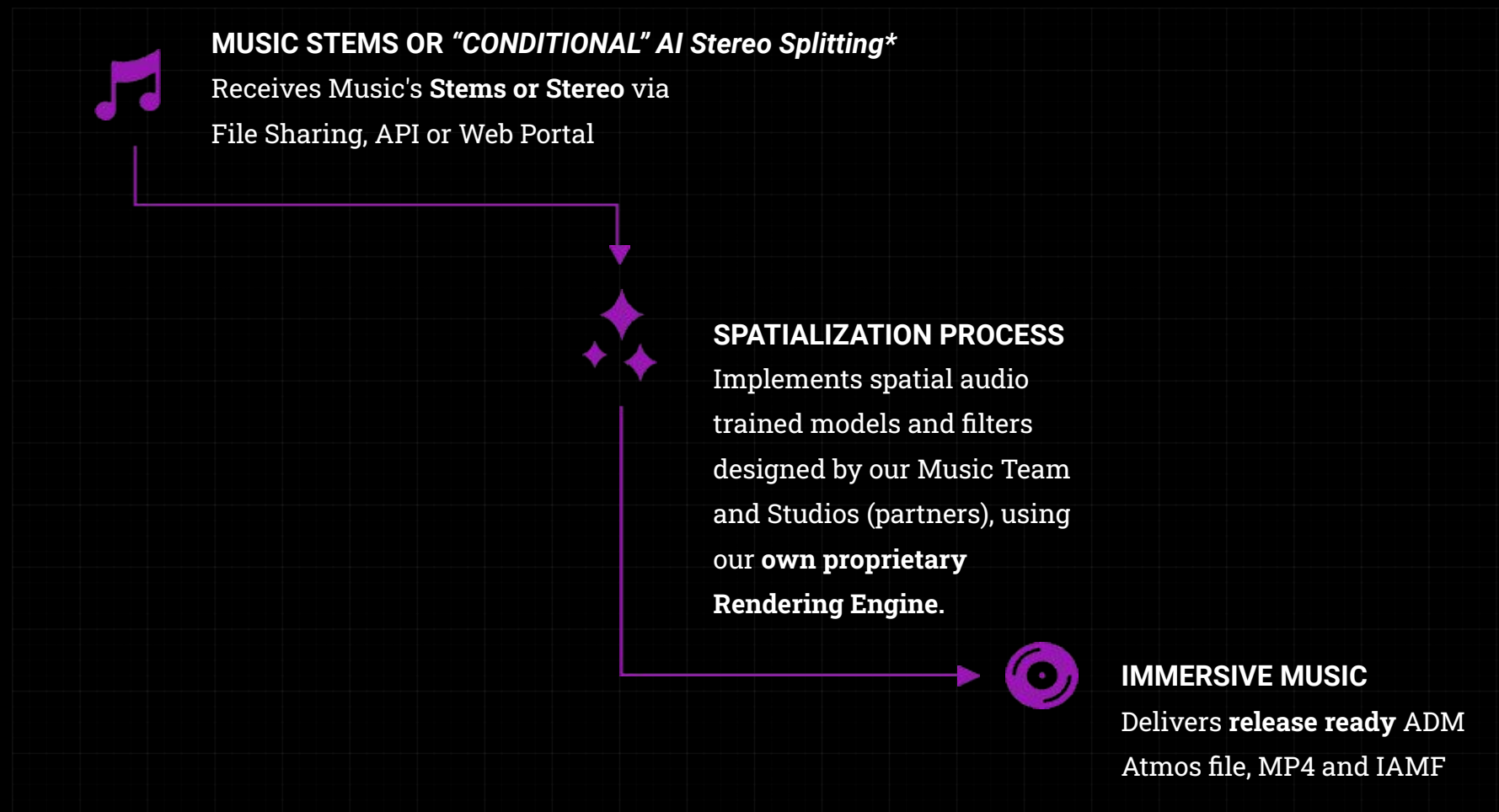
SPATIAL9 IS POISED TO REDEFINE THE WAY PEOPLE EXPERIENCE SOUND, UNLOCKING NEW CREATIVE POSSIBILITIES AND IMMERSING LISTENERS IN UNPRECEDENTED AUDIO DIMENSIONS AND SCALE.

MISSION: To elevate the audio experience of every person and foster deeper emotional connections through sound.

VISION: To deliver the most powerful immersive audio platform of the century by revolutionizing the sound experience.



Our Technology



*Conditional stereo-to-STEM AI splitting is used when traditional STEM files are unavailable due to issues such as difficult artifact recovery, crosstalk or noisy effects from older recordings, or damaged media. This method is particularly relevant for processing old music collections or real-time stereo Music related to Videos uploads.





REST API

- ❖ **ARCHITECTING WEB SERVICES FOR EFFORTLESS DATA EXCHANGE AND INTEROPERABILITY**
- ❖ **SYSTEMS COMMUNICATION WITH HTTP METHODS SUCH AS GET, POST, PUT AND DELETE**
- ❖ **STATELESSNESS**
- ❖ **ENABLES SCALABILITY AND RELIABILITY**
- ❖ **BRIDGE BETWEEN APPLICATION AND EXTERNAL SYSTEMS**
- ❖ **EXECUTION OCCURS SEAMLESSLY, ENSURING TRANSPARENCY TO THE USER**
- ❖ **ALLOWS EASY INTEGRATION WITH NEW PARTNERS**
- ❖ **ENHANCED SECURITY: APPLICATION CORE ISOLATED FROM THE USER AND STRONG AUTHENTICATION**



- ❖ **WE HAVE OUR OWN RENDERING ENGINE**
- ❖ **DOLBY-VAILD METADATA**
- ❖ **APPLIES OBJECT-BASED MOVEMENTS**
- ❖ **SUPPORTS FOR CHANNEL-BASED**
- ❖ **SUPPORTS BINAURAL METADATA / DOLBY HEADPHONE**



- ❖ **TRANSLATE ADM ATMOS METADATA TO IAMF METADATA**

- ❖ *For ADM files generated into our Rendering Engine, roadmap includes “**any** ADM file”*

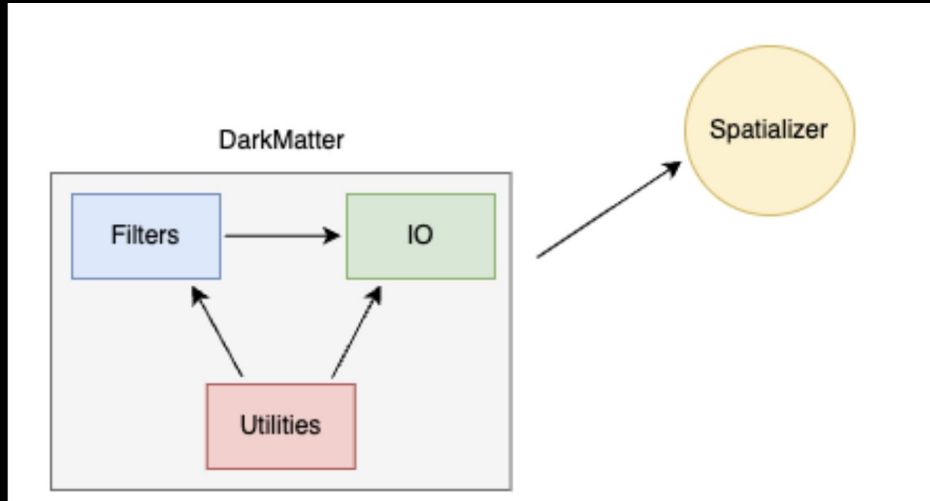
- ❖ **SUPPORTS CHANNEL-BASED**

- ❖ **SUPPORTS SCENE-BASED WITH AMBISONICS**

- ❖ **AMBISONICS DYNAMICS AND MOVEMENTS**

- ❖ **SUPPORTS BINAURAL HEADPHONE RENDERING MODE**

DarkMatter



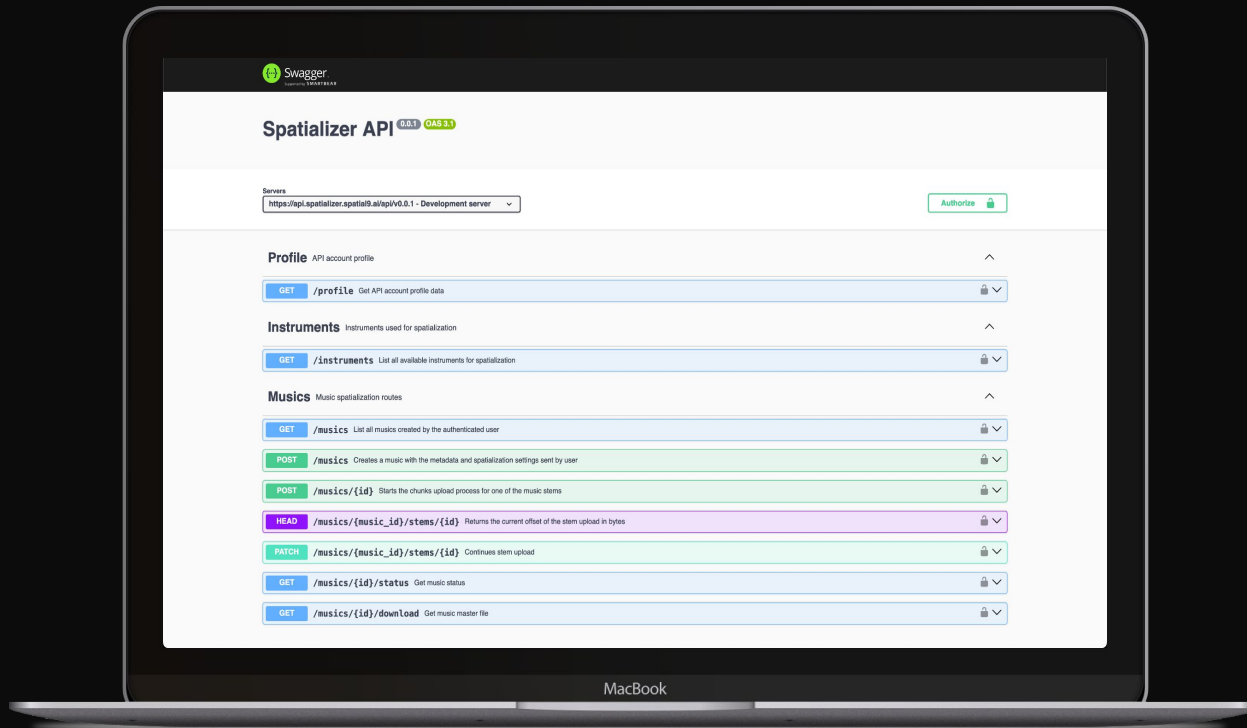
★ SPATIAL9 CORE LIBRARY

★ FUNDAMENTAL TOOLS FOR AUDIO DSP

- Utilities - Common used math functions for DSP
- Filters - Signal processors (HPF, LPF, Reverb)
- IO - Audio formats readers and writers (RIFFWave, Atmos, IAMF, MP4)

★ CODE OPTIMIZATION, REUSE AND COMPANY KNOWLEDGE BASE

Spatializer API - For Direct Partner Integration



★ SEAMLESS COLLABORATION

- Connect with partners effortlessly, enabling smooth data exchange and interaction!

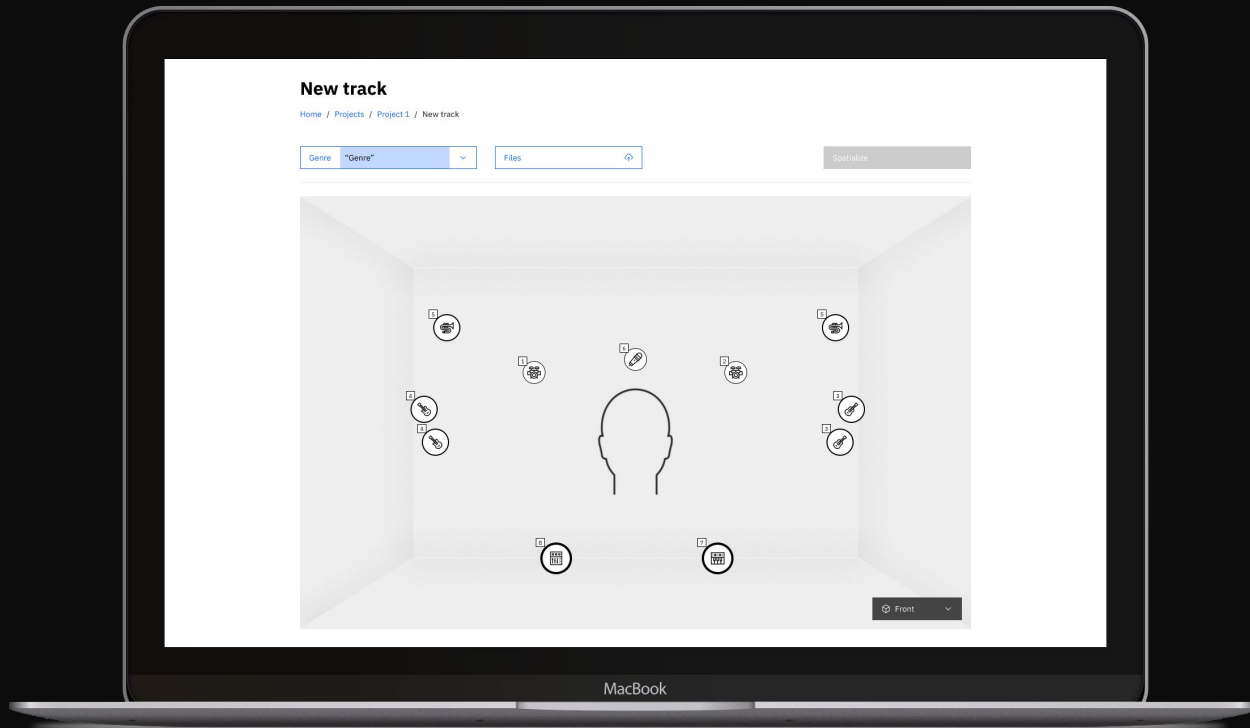
★ CUSTOMIZABLE INTEGRATION

- Easily tailor the API to fit your unique business needs, making integration a breeze!

★ ROBUST SECURITY MEASURES

- Ensure your data is protected with strong authentication and isolated application cores, keeping your integrations safe!

Portal - For Studios and Partners



★ END USER PORTAL

- Do it yourself, effortlessly!

★ EASILY MAKE IT IMMERSIVE

- Drag and drop your Music Stems, let the magic happen!

★ FULL COMPLIANCE

- The technical requirements for distribution and streaming have been satisfied, it's release ready!

Simplify
Spatial
Sound



SPATIAL⁹

Machine Learning for multichannel audio format conversion?

Analogies – recent successful developments: stem/source separation, dereverberation.

Differences – spatial vs. temporal analysis; need for multichannel signal references.

Some new challenges – defining appropriate metrics and loss functions; collecting representative data.

Some opportunities – audio+image presentation: spatial coincidence is a more significant consideration in today's industry

- audio/video and media coding
- multimodal machine learning techniques
- immersive mastering assistant.