# Leveraging Frequency Based Salient Spatial Sound Localization to Improve 360° Video Saliency Prediction

Mert Cokelek
Hacettepe University
mert.cokelek0699@gmail.com

Nevrez Imamoglu
AIST Japan
nevrez.imamoglu@aist.go.jp

Cagri Ozcinar
Samsung Electronics
cagriozcinar@gmail.com

Erkut Erdem
Hacettepe University
erkut@cs.hacettepe.edu.tr

Aykut Erdem
Koç University
aerdem@ku.edu.tr

## Abstract

*Virtual and augmented reality (VR/AR) systems have dramatically gained in popularity with various application areas such as gaming, social media, and communication. It is therefore a crucial task to have the know-how to efficiently utilize, store or deliver 360° videos for end-users. Towards this aim, researchers have been developing deep neural network models for 360° multimedia processing and computer vision fields. In this line of work, an important research direction is to build models that can learn and predict the observers' attention on 360° videos to obtain so-called saliency maps computationally. Although there are a few saliency models proposed for this purpose, these models generally consider only visual cues in video frames by neglecting audio cues from sound sources. In this study, an unsupervised frequency-based saliency model is presented for predicting the strength and location of saliency in spatial audio. The prediction of salient audio cues is then used as audio bias on the video saliency predictions of state-of-the-art models. Our experiments yield promising results and show that integrating the proposed spatial audio bias into the existing video saliency models consistently improves their performances.*

## 1 Introduction

Several studies [1–3], have shown that auditory inputs influence human attention mechanism, yet many visual saliency estimation models neglect the effects of auditory cues when estimating saliency maps. In addition, from a computational perspective, some of the earlier works [4–8] also demonstrated that visually salient cues are partly correlated with the audio source location and semantics. In particular, Tavakoli *et al.* [4] proposed an audio-visual deep learning model (DAVE), which has an encoder-decoder architecture for video saliency prediction. Min *et al.* [5] suggested a novel multimodal saliency (MMS) model for audio-visual attention, which is proposed to be combined with the existing deep learning-based saliency models with a late-fusion, and to promote their performance by an average of 5%. Tsiami *et al.* [6] proposed a single multimodal network (STAViS) for audio-visual saliency, which learns to localize sound sources and to fuse the audio and visual saliency maps. Chen *et al.* [7] proposed a deep neural network architecture for feature extraction, semantic interaction, and their fusion for auditory and visual inputs.

On the other hand, to the best of our knowledge, *spatial audio* information has just begun to be used for saliency prediction in 360° videos [9]. Moreover, with the advent of 360° videos, the saliency prediction task has faced new challenges. Humans do not discover their 360° environments at a glance, but starting within a narrower viewport, and then they continuously work out the peripheries with their head/eye movements. In contrast, 360° video sequences are represented as fully observable to computers. This results in a contradictory behavior for perception between computers and humans. At this point, spatial audio cues, which provide directional information in 360° space can be incorporated alongside the visual cues to supply additional insight for localizing the saliency predictions in 360° videos.

In this paper, the 360° video saliency prediction task is addressed by leveraging the spatial audio information to localize the audio saliency and enhance the output of the existing (audio-)visual saliency prediction models. For this purpose, we have adapted the mel-cepstrum based spectral residual saliency detection model (MCSR) proposed by Imamoglu *et al.* [10], to spatial audio. Despite it was presented as an image saliency model, it is highly adaptable to audio processing since it consists of mel-frequency cepstral coefficients (MFCCs). The proposed audio saliency localization model is built for first-order ambisonics (FOA) in 4-channel B-format, which is widely used in VR applications for spatial audio experience. Each channel $(W, X, Y, Z)$ in FOA-encoded audio represents a different directionality in the 360° space: center, forward-

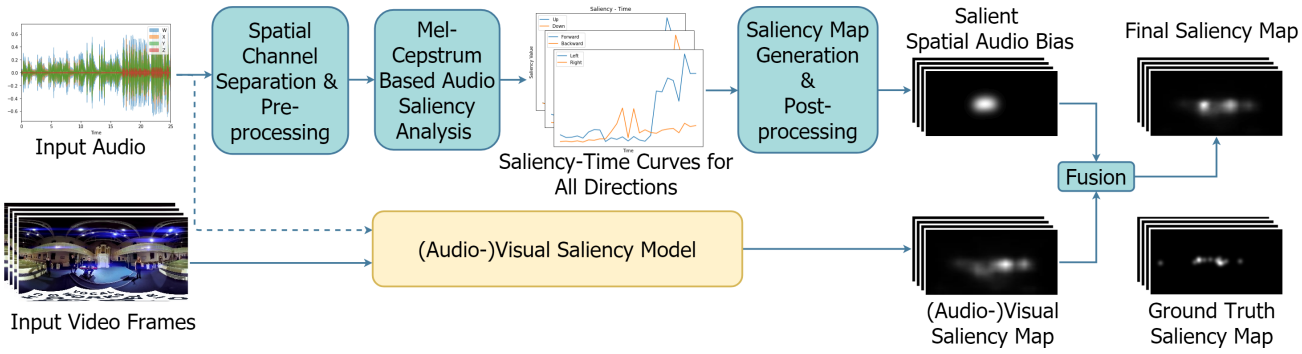


Figure 1. System overview of the proposed approach.

backward, left-right and up-down, respectively [11]. The MCSR model is adapted to FOA by applying on each channel to localize the salient sound in 360° space, and also detect the strength of audio saliency in the time domain. The produced audio saliency maps are further combined with the existing five video saliency models with a late-fusion. **The contributions of our work can be summarized as follows:**

- We show that MFCC-based signal analysis can provide information about audio saliency in the time domain. This can be used in any audio saliency model to improve its precision by weighting (suppressing/boosting) the audio predictions according to their saliency values for any instant.
- By extending the MFCC analysis for FOA, we show that it is possible to localize salient sounds in 360° audio.
- Combining the first two findings, we show that integrating this lightweight unsupervised salient spatial sound localization (SSSL) method as a bias to the existing traditional/deep learning-based (audio-)visual saliency models can improve their performance by an average of 14%, on 360° videos.

The rest of the paper is organized as follows: Section 2 describes our method. The dataset, experimental details, performance analysis, and comparison with the state-of-the-art models are given in Section 3. Section 4 covers the conclusion and future work.

## 2 Method

The framework of our proposed audio saliency model is illustrated in Figure 1. The aim is to find the exact location and strength of salient spatial audio in the time domain, as a bias for the video saliency prediction models. Firstly, the FOA waveforms have been pre-processed for the MFCC-based saliency analysis. Then, the saliency analysis has been performed on each direction independently and the results have been combined in a 3D space. Lastly, spatial audio saliency maps have been produced and fused with the outputs of the existing state-of-the-art (audio-)visual saliency prediction models.

**Preprocessing.** The audio waveforms have been divided into shorter clips. By considering shorter clips, the model is expected to highlight the locally and perceptually important sounds better. Otherwise, the entire audio is given to the model and hence, the local sounds may be overlooked. In our experiments, the effects of different clip durations from half second to original video length have been investigated. Additionally, the performance of multi-scale saliency analysis in the time domain has been observed to see the effect of local and global perception of sound together. In this approach, clips of different durations have been analyzed independently and combined in the time domain (such that for a video, four Saliency-Time curves have been obtained from clip duration of (1) one second, (2) quarter of the total duration, (3) half of the total duration, (4) total duration.). The best results have been obtained with one second-clips, without a multi-scale approach. Clip durations shorter than one second have resulted in dense peaks and longer ones have resulted in invariability to saliency in the time domain. In the rest of the paper, the experimental analysis and performance comparison are based on the pre-processing of audio into one second-clips. Finally, to localize the saliency in the 360° space, the FOA waveforms have been decomposed to six directions (forward, backward, left, right, up, down) by adapting the decoding formulas in [11] as:

$$\begin{aligned} P &= (\sqrt{2}W + C) * 2 \\ N &= (\sqrt{2}W - C) * 2 \end{aligned} \tag{1}$$

where $W$ is the center channel and $P$, $N$ correspond to positive and negative directions for a given channel $C$, respectively.

**Extending the MCSR model for 360° audio saliency prediction.** The MCSR saliency model is applied to each channel independently. As shown in

Figure 2, the waveforms corresponding to the positive and negative directions are given to the MCSR model separately for saliency localization in one channel. The resulting Saliency-Time curves are then subtracted from each other and normalized into $[0, 1]$ range. For a given time $t$, a saliency value above 0.5 implies that the salient sound for that channel is in the positive direction, and vice versa. By applying this procedure on all channels, we obtain a 3D vector for each audio sample, representing the direction of arrival of the salient sound on the 360° space.

**Saliency map generation and post-processing.** The sampling rate of the audios and the frame rates of the videos are not equal, thus the audio predictions are framed to have a one-to-one correspondence with the video frames. The obtained 3D vectors for a given audio frame are first converted to unit vectors, then transformed to (u,v) image coordinates and the corresponding pixels in the output audio saliency map have been highlighted. Then, a Gaussian filter with a kernel size of $h/2$, where $h$ denotes the height of the map, is applied to the resulting attentive sound location or fixation points. This procedure results in an auditory saliency map, which is finally scaled to $[0, 1]$ range. The produced saliency maps refer to the most salient sounds per frame in the 360° videos. However, they do not provide information about the strength of saliency. In MMS [5], the audio saliency predictions have been weighted by their reliability before the fusion. Motivated by this idea, we have weighted the localized saliency predictions based on the Saliency-Time curves obtained by applying the MCSR saliency model on channel $W$. This weighting operation is done in a linear fashion, by normalizing the saliency values in the curve into $[0, 1]$ range and multiplying every audio saliency map with the saliency value of that instant. If the audio is worth attracting human attention, the corresponding audio saliency map will have a higher energy, and vice versa. Finally, we have utilized the temporal information in audio saliency maps to obtain smoother and more natural transitions between the predictions. In our experiments, the best results are obtained by averaging the last $n$ audio frames where $n$ denotes the fps value of the video.

**Audio-visual saliency fusion.** As illustrated in Figure 1, the final step is to integrate the output of the audio saliency predictions with the existing (audio-)visual saliency models to generate the final prediction as:

$$S = f(S_a, S_v), \tag{2}$$

where $S$ is the final audio-visual saliency map, $S_a$ is the output of our audio saliency model, $S_v$ is the output of an existing (audio-)visual saliency model, and $f$ is the fusion operation. We have used an integration scheme inspired by Itti-Koch's quantization and averaging based fusion strategy [12, 13], as given below:

$$f(S_a, S_v) = 0.5 * N(S_a, M) + 0.5 * N(S_v, M), \tag{3}$$

where $N$ represents the quantizer which transforms the saliency map into $M$ discrete levels. In our experiments, $M$ is empirically selected as 8.

## 3 Experiments

**Dataset.** For testing, we have used the dataset provided by Chao *et al.* [8] which contains 12 omnidirectional videos (ODVs) with FOA in 4-channel $B$-format with a duration of 25 seconds. Each ODV has been split into three categories: Conversation, Music, and Environment. The fixation points have been collected from a total number of 45 subjects, where each video has been viewed by randomly selected 15 subjects and each subject has viewed each ODV once. The frame rate of videos varies from 25 fps to 60 fps and the audio sampling rate of all videos is 48000 Hz.

**Comparison with the state-of-the-art.** For performance criteria, we have employed the five commonly used saliency evaluation metrics [14]: AUC-Judd, NSS, CC, SIM, and KL Divergence. Table 1 shows the performance comparison of five state-of-the-art models with and without the proposed spatial audio saliency fusion. For additional analysis, the performance of fusion with the SSSL maps is compared with that of the audio energy maps (AEMs). Audio energy maps for FOA represent the direction of arrival of the 360° audio, and we have obtained them by using the strategy in [15]. For comparison, we have chosen CP360 [16] as a visual saliency model for 360° videos, MMS and STAViS as audio-visual saliency models and UNISAL [17] as a visual saliency model for 2D videos. Lastly, we have considered the recently proposed AVS360 model [9], which performs 360° audio-visual saliency prediction as an upper bound for the evaluated models as it is trained on this dataset. To fuse audio information, this model employs the extracted multi-channel AEMs. Inspired by the evaluation in [9], we have included an equator bias to every predicted saliency map before performing the quantitative analysis on the videos. As shown in Table 1, the proposed spatial audio saliency fusion when applied to the results of the existing models gives rise to better predictions in terms of all metrics other than AUC-J. For instance, we observe on average 24.9% and 25.5% performance gains for NSS and CC metrics, respectively. It is important to note that, as mentioned in [14], these metrics are considered as the most reliable evaluation metrics for saliency prediction on capturing the viewing behaviors. In Figure 3, we also present some sample qualitative comparisons. As seen from these sample frames, the proposed spatial audio-driven post-processing better localizes the salient regions in 360° videos and eliminates the false positives for almost all of the samples. In the table, it can also be seen that the SSSL results are better than the AEM fusion results, which motivates us to build novel architectures for SSSL fusion to outperform the existing state-of-the-art.

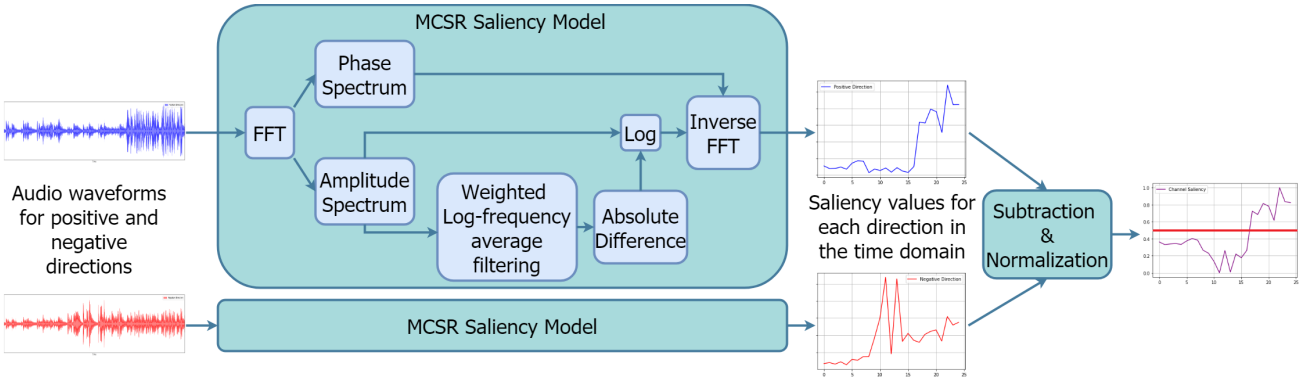


Figure 2. Overview of the MCSR Saliency Model and salient spatial sound localization for one audio channel.

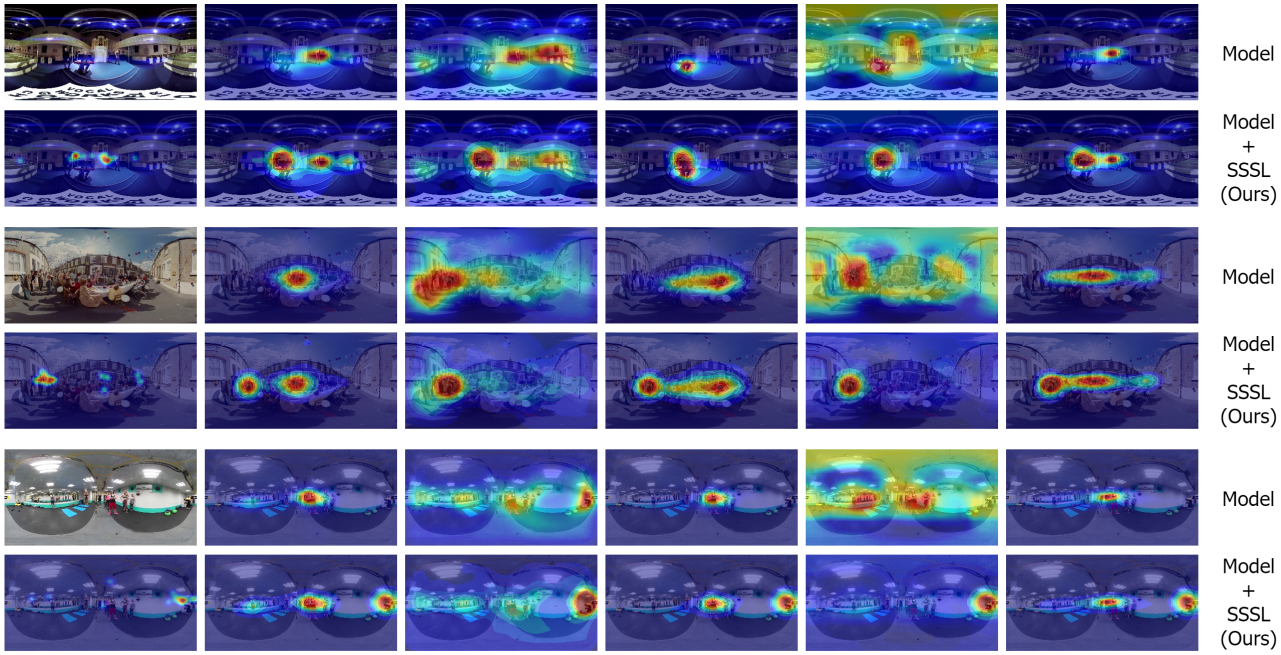


Figure 3. Qualitative evaluation of the (audio-)visual saliency models with the proposed audio saliency fusion.

## 4 Conclusion

In this paper, we have investigated the effect of spatial audio cues alongside visual cues for 360° video saliency prediction. We have proposed a spatial audio saliency prediction model for localizing the salient sounds in 360° space, finding the strength of saliencies in the time domain, and producing audio saliency maps for late-fusion to any (audio-)visual saliency model. The results demonstrate that the proposed SSSL model has more contribution than AEMs and the other audio saliency models to 360° saliency prediction task, and their fusion to the existing (audio-)visual saliency models improve their performance by an average of 14%. Developing novel neural architectures for salient spatial sound localization & fusion for 360° videos is left as future research.

## Acknowledgements

This work was supported in part by GEBIP 2018 Award of the Turkish Academy of Sciences to E. Erdem, BAGEP 2021 Award of the Science Academy to A. Erdem, and by TUBITAK-1001 Program Award No. 120E501.

## References

[1] J. Vroomen and B. d. Gelder, "Sound enhances visual perception: cross-modal effects of auditory orga-

Table 1. Performance comparison with the state-of-the-art saliency models.

| Model | AUC-J↑ | NSS↑ | CC↑ | SIM↑ | KL↓ |
|---|---|---|---|---|---|
| STAViS | **0.839** | 1.938 | 0.346 | 0.259 | 14.43 |
| STAViS + AEM | 0.721 | 1.925 | 0.347 | 0.258 | 14.52 |
| STAViS + SSSL (Ours) | 0.740 | **2.067** | **0.374** | **0.279** | **13.91** |
| Perf. Gain | ⇓ -11.7% | ⇑ 6.6% | ⇑ 8.1% | ⇑ 7.7% | ⇑ 1.8% |
| MMS | **0.837** | 1.347 | 0.250 | 0.171 | 18.15 |
| MMS + AEM | 0.819 | 1.471 | 0.272 | 0.187 | 17.66 |
| MMS + SSSL (Ours) | 0.827 | **1.579** | **0.293** | **0.195** | **17.33** |
| Perf. Gain | ⇓ -1.2% | ⇑ 17.2% | ⇑ 17.2% | ⇑ 14.0% | ⇑ 4.7% |
| UNISAL | **0.804** | 1.288 | 0.234 | 0.224 | 15.20 |
| UNISAL + AEM | 0.576 | 1.361 | 0.252 | 0.227 | 14.95 |
| UNISAL + SSSL (Ours) | 0.740 | **1.939** | **0.360** | **0.297** | **12.61** |
| Perf. Gain | ⇓ -7.9% | ⇑ 50.5% | ⇑ 53.8% | ⇑ 22.2% | ⇑ 20.5% |
| CP360 | **0.842** | 1.165 | 0.224 | 0.144 | 19.40 |
| CP360 + AEM | 0.831 | 1.369 | 0.257 | 0.169 | 18.13 |
| CP360 + SSSL (Ours) | 0.834 | **1.462** | **0.276** | **0.179** | **17.71** |
| Perf. Gain | ⇓ -0.9% | ⇑ 25.5% | ⇑ 23.2% | ⇑ 24.3% | ⇑ 9.5% |
| Avg Perf. Gain | ⇓ -5.4% | ⇑ 24.9% | ⇑ 25.5% | ⇑ 17.0% | ⇑ 9.1% |
| AVS360 | **0.769** | **2.656** | **0.453** | **0.349** | **10.27** |


nization on vision.," *Journal of experimental psychology: Human perception and performance*, vol. 26, no. 5, p. 1583, 2000.

[2] Y. Chen, T. V. Nguyen, M. Kankanhalli, J. Yuan, S. Yan, and M. Wang, "Audio matters in visual attention," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 11, pp. 1992–2003, 2014.

[3] X. Min, G. Zhai, Z. Gao, C. Hu, and X. Yang, "Sound influences visual attention discriminately in videos," in *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 153–158, IEEE, 2014.

[4] H. R. Tavakoli, A. Borji, E. Rahtu, and J. Kannala, "Dave: A deep audio-visual embedding for dynamic saliency prediction," *arXiv preprint arXiv:1905.10693*, 2019.

[5] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Transactions on Image Processing*, vol. 29, pp. 3805–3819, 2020.

[6] A. Tsiami, P. Koutras, and P. Maragos, "Stavis: Spatio-temporal audiovisual saliency network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4766–4776, 2020.

[7] J. Chen, Q. Li, H. Ling, D. Ren, and P. Duan, "Audiovisual saliency prediction via deep learning," *Neurocomputing*, vol. 428, pp. 248–258, 2021.

[8] F.-Y. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Audio-visual perception of omnidirectional video for virtual reality applications," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, IEEE, 2020.

[9] F.-Y. Chao, C. Ozcinar, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Towards audio-visual saliency prediction for omnidirectional video with spatial audio," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 355–358, IEEE, 2020.

[10] N. İmamoğlu, Y. Fang, W. Yu, and W. Lin, "2d melcepstrum based saliency detection," in *2013 IEEE International Conference on Image Processing*, pp. 236–239, IEEE, 2013.

[11] D. Arteaga, "Introduction to ambisonics," *Escola Superior Politècnica Universitat Pompeu Fabra, Barcelona, Spain*, 2015.

[12] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10-12, pp. 1489–1506, 2000.

[13] J. Bauer, "Simple itty-koch-style saliency maps." `gist.github.com/tatome/d491c8b1ec5ed8d4744c`, 2016.

[14] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[15] P. Morgado, N. Vasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360 video," *arXiv preprint arXiv:1809.02587*, 2018.

[16] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun, "Cube padding for weakly-supervised saliency prediction in 360 videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1420–1429, 2018.

[17] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *European Conference on Computer Vision*, pp. 419–435, Springer, 2020.