Deep Audio-Visual Beamforming for Speaker Localization

Xinyuan Qian, *Member, IEEE*, Qiquan Zhang, *Member, IEEE*, Guohui Guan, and Wei Xue*, Member, IEEE*

***Abstract*—Generalized Cross Correlation (GCC) is the most popular localization technique over the past decades and can be extended with the beamforming method e.g. Steered Response Power (SRP) when multiple microphone pairs exist. Considering the promising results of Deep Learning (DL) strategies over classical approaches, in this work, instead of directly using Generalized Cross Correlation (GCC), SRP is derived with the DL-learnt ideal correlation functions for each pair of a microphone array. To deploy visual information, we explore the Conditional Variational Auto-Encoder (CVAE) framework in which the audio generative process is conditioned on the visual features encoded by face detections. The vision-derived auxiliary correlation function eventually contributes to the back-end beamformer for improved localization performance. To the best of our knowledge, this is the first deep-generative audiovisual method for speaker localization. Experimental results demonstrate our superior performance over other competitive methods, especially when the speech signal is corrupted by noise.**

***Index Terms*—Audio-visual fusion, speaker localization, variational auto-encoder.**

I. Introduction

Sound source localization has been a long-standing research problem over the past decades and is a crucial task in various audio applications, such as automatic camera steering, speech enhancement [1], and speaker diarization [2].

By exploiting the acoustic signal's characteristics, classical Sound Source Localization (SSL) methods rely on mathematical models and various optimization schemes to locate a sounding object. Among different approaches, Time Difference of Arrival (TDoA)-based methods account for the vast majority, where GCC, the building block of most SSL algorithms [3], estimates the sound location at the time delay which maximizes the cross-correlation function. To compromise between the sharpness of a GCC peak at the ground truth and its sensitivity at errors, several weighting functions are examined [4] where the most prevalent is the one with the PHAse Transform (PHAT) function i.e., Generalized Cross Correlation with Phase Transform (GCC-PHAT).

Beamforming methods estimate the acoustic power [5] on a predefined grid of potential sound locations in multi-microphone scenarios. In particular, SRP has the highest popularity due to its simplicity and high performance [6], [7]. By consolidating the multi-channel GCC, SRP steers at every grid point and locates the sound with the maximum SRP value. To counteract the multi-path effect and room reverberation, GCC is replaced by GCC-PHAT with the resulting method named Steered Response Power PHAse Transform (SRP-PHAT) [8], [9]. Although the performance in noisy and reverberant environments has been greatly improved by exploiting multiple pairwise correlations [9], it is still far from satisfactory.

SSL can also be tackled through a learning process. For example, in [10], Support Vector Machine (SVM) vectors are used as subspaces for acoustic space mapping. Recently, assisted by large-scale annotated data, DL models superior classical approaches by directly mapping input acoustic features to sound locations. Typical work includes Direction of Arrival (DoA) prediction from time-frequency features, i.e., Short-Time-Fourier-Transform (STFT) [11] and time-delay features, i.e., GCC-PHAT [12], [13], given a back-end location classifier. Other alternative methods may infer DoA from the eigenvectors of the MUltiple SIgnal Classification (MUSIC) algorithm or solve the localization problem in an end-to-end manner [14]. Moreover, enerative models [15], [16] have recently been explored with mis-matched array geometry conditions and limited training data, respectively. The stochastic neural networks involved with the encoder-decoder scheme lead to promising results.

Unlike audio signals, vision is not corrupted by acoustic perturbations; therefore, it can contribute to SSL, especially under low Signal-to-Noise Ratio (SNR) conditions. On the contrary, vision is penalized by clutters and varying lighting conditions, and cannot perform when target moves outside the camera's field of view. Thus, audio and vision have modality-specific limitations and complementary characteristics where their fusion benefits have been proved by several studies. For example, a sounding object can be precisely localized either on a 2D [17], [18] or a 360° image plane [19]. Despite the impressive success, audio-visual SSL has been much less investigated than its audio-only counterparts [12], [15]. Significant improvement has been observed in a recent work [20] in which complementary vision features are concatenated as inputs to the back-end sound

Manuscript received February 28, 2022; revised March 27, 2022; accepted April 3, 2022. Date of publication April 6, 2022; date of current version May 10, 2022. This work was supported by the Science and Engineering Research Council, Agency of Science, Technology and Research (A * STAR), Singapore, through the National Robotics Program under Grant 192 25 00054. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yu Tsao. *(Corresponding author: Qiquan Zhang.)*

Xinyuan Qian, Qiquan Zhang, and Guohui Guan are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: eleqian@nus.edu.sg; zhangqiquan_hit@163.com; e0516157@u.nus.edu).

Wei Xue is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: weixue@comp.hkbu.edu.hk).

Digital Object Identifier 10.1109/LSP.2022.3165466

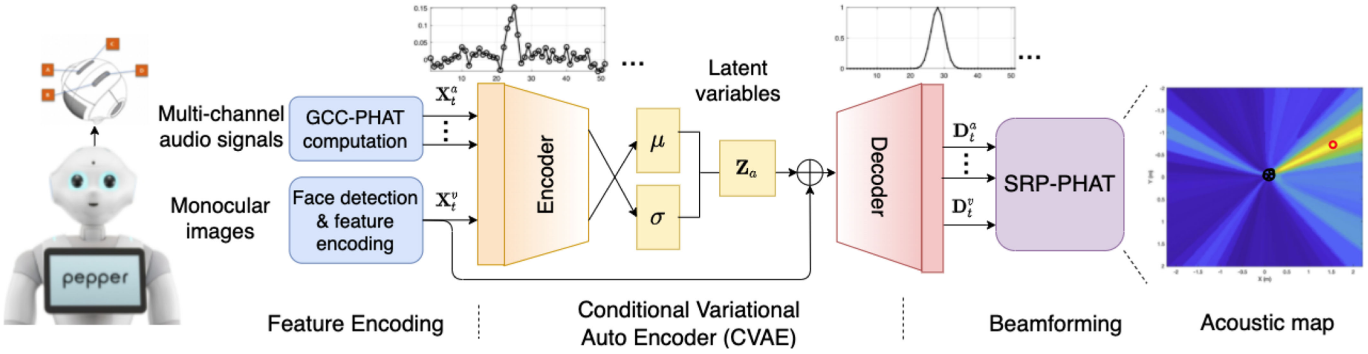


Fig. 1. The proposed Deep Generative Beamformer (DGB) architecture for audio-visual speaker localization. Given the received multi-channel GCC-PHATs and the extracted visual features, CVAE encodes the latent variables and decodes multiple Gaussian-shaped correlation functions which peak at the ground truth time delays. The reconstructed signals are then used for the back-end beamforming (i.e., SRP-PHAT) to generate the acoustic map (yellow and blue indicate the higher and lower probability of including a sound source (red circle), respectively.). $\oplus$ stands for concatenation.

location classifier. Despite this, how to incorporate vision into a generative localization model remains an unexplored problem.

In this paper, we propose to combine the beamforming technique with a DL procedure for speaker localization. To our best knowledge, we are the first deep generative audio-visual SSL method and our contributions are summarized as follows:

1) We design a Variational Autoencode (VAE) framework which encodes the multi-channel GCC-PHATs to reconstruct multiple Gaussian-shaped correlation functions that peak at the real time delay of each microphone pair.
2) For visual incorporation, we use CVAE to fuse image features as the auxiliary inputs and model the camera as a virtual microphone pair which provides a correlation function to the back-end SRP-PHAT beamformer.
3) We deploy the denoising capability of VAE by feeding the proposed method with noisy GCC-PHATs at various SNRs. This results in an improved noise robustness.

II. Proposed Method

Given the synchronized multi-channel audio signals $\mathbf{s}_t^{1:N}$ and monocular image $\mathbf{I}_t$ at time t, with known microphone locations $\mathbf{o}^{1:N}$, our objective is to find the instant speaker DoA (i.e. azimuth) towards the sensing platform, formulated as:

$$\hat{\theta}_t = \mathcal{F}(\mathbf{s}_t^{1:N}, \mathbf{I}_t, \mathbf{o}^{1:N}), \tag{1}$$

where $\mathcal{F}$ is the proposed system, N is the total number of microphones, and $\hat{\theta}_t$ is the estimate of the speaker DoA θ_t.

To this end, we propose the Deep Generative Beamformer (DGB) which incorporates a DL process with the conventional SRP-PHAT beamforming for audio-visual SSL. Fig. 1 illustrates the system architecture, which mainly consists of three stages: (1) the front-end feature encoding; (2) the DL-based VAE, and (3) the Signal Processing (SP)-based back-end beamforming (for acoustic map generation). It should be noted that the DL parameters are only optimized at VAE. In addition, we also propose a contrastive method with audio-only inputs, denoted A-DGB, where VAE replaces CVAE without vision contribution.

We elaborate each block in next subsections. For brevity, the time index t is eliminated in Sections II-A and II-D.

A. Feature Encoding

1) Audio Processing: SSL studies can be grouped into time-delay methods [21], [22], energy ratio methods [23], [24], and machine learning methods [25], [26]. Among them, time-delay methods are widely adopted due to their effectiveness. In particular, GCC-PHAT, which facilitates the TDoA estimation between any two arbitrary microphones [12], is more robust to noise and room reverberations [27]. It peaks at the actual time delay of a sounding object, and herein, we use it as the acoustic feature. Let $\mathbf{S}_{n_1}$ and $\mathbf{S}_{n_2}$ $\{(n_1, n_2), \forall n_1 < n_2 \leq N\}$ denote the STFT of short-time audio signals received by a microphone pair. We use $m \leq M$ for the index with M as the total pair number. Then, GCC-PHAT is computed as:

$$g^m(\tau) = \sum_k \mathcal{R}\left(\frac{\mathbf{S}_{n_1}(k)\mathbf{S}_{n_2}^*(k)}{|\mathbf{S}_{n_1}(k)\mathbf{S}_{n_2}^*(k)|} e^{i\frac{2\pi k}{N_s}\tau}\right), \tag{2}$$

where $\tau \in [-\frac{L}{2}, \frac{L}{2}]$ is the time delay with L the maximum allowable length, i indicates the imaginary unit, $*$ denotes the complex conjugate, k and $\mathcal{R}$ indicates the frequency bin and the real part of a complex number, respectively, and N_s is the STFT length. We concatenate all the GCC-PHATs, denoted as $\mathbf{X}^a = \oplus g^{m=1\ldots M} \in \mathcal{R}^{1\times ML}$, as the input acoustic feature.

2) Video Processing: We encode the visual image feature with a multi-variant Gaussian distribution as proposed in [20]. Given the detected face bounding box $\mathbf{b} = (u, v, w, h)^\intercal$, we extract the face central point as $\boldsymbol{\mu} = (u + \frac{1}{2}w, v + \frac{1}{2}h)^\intercal$. The encoded multi-variant Gaussian distribution (in u and v directions) centralized at $\boldsymbol{\mu}$ is given as:

$$\mathcal{V}(\mathbf{x}) = \begin{cases} \max_d \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^\intercal} & \mathbf{b} \neq \emptyset, \\ \mathcal{U}(\mathbf{x}) & \text{otherwise}, \end{cases} \tag{3}$$

where $\mathbf{x}$ is an arbitrary 2D image location, $\Sigma = diag(w_d^2, h_d^2)$ is a diagonal covariance matrix, and $\mathcal{U}(\mathbf{x})$ indicates a uniform distribution. We re-scale and concatenate the horizontal and vertical components of $\mathcal{V}(\mathbf{x})$ and $\mathbf{X}^v \in \mathcal{R}^{1\times ML}$.

B. A-DGB: VAE With Audio Inputs

We perform acousitc SSL with the proposed A-DGB method. The encoder approximates the intractable posterior distribution

TABLE I
EXPERIMENTAL RESULTS REPORTED IN MAE (↑) AND ACC (↓) IN THE SSLR HUMAN TEST SET [31]

	Ref	MAE(°) ↓	ACC (%) ↑
Video	Face-MLP	16.85	32.2
Audio	SRP-PHAT [8]	2.92	88.1
	GCC-MLP [31]	4.75	95.1
	deepGCC [15]	3.40	95.3
	A-DGB (ours)	3.12	96.4
Audio-visual	AVMLP [20]	1.84	97.1
	DGB (ours)	**1.84**	**98.0**

$Q_\phi(\mathbf{Z}|\mathbf{X}_t^a)$ with the network parameters ϕ, while the decoder generates the conditional distribution $P_\gamma(\mathbf{X}_t^a|\mathbf{Z})$ with the parameters γ. Specifically, the multi-channel GCC-PHAT $\mathbf{X}_t^a$ are encoded in fixed-size latent variables $\mathbf{Z}_t$ while the decoder reconstructs multiple Gaussian-like functions $\mathbf{D}_t^a$. For the m-indexed microphone pair, the output of A-DGB should peak at the real $\tilde{\tau}^m$:

$$d^m(\tau) \sim exp\left(-\frac{|\tau-\tilde{\tau}^m|^2}{2\sigma_D^2}\right), \tag{4}$$

where σ_D is the standard deviation, and the entire output is the concatenated d^m, i.e., $\mathbf{D}_t^a = \oplus\, d^{m=1\ldots M} \in \mathcal{R}^{1\times ML}$.

According to variational inference [29], [30], VAE aims to make the encoder distribution $Q_\phi(\mathbf{Z}_t|\mathbf{X}_t^a)$ consistent with $P_\gamma(\mathbf{Z}_t|\mathbf{X}_t^a) \propto P_\gamma(\mathbf{X}_t^a|\mathbf{Z}_t)P(\mathbf{Z}_t)$, which can be achieved by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}^a = \mathbb{E}\left(logP_\gamma(\mathbf{X}_t^a|\mathbf{Z}_t)\right) - KL\left(Q_\phi(\mathbf{Z}_t|\mathbf{X}_t^a)|P(\mathbf{Z}_t)\right), \tag{5}$$

where the first item is interpreted as the encoder's reconstruction error and the second item (negative KL divergence between $Q_\gamma(\mathbf{Z}_t|\mathbf{X}_t^a)$ and the prior probability $p(\mathbf{Z}_t) = \mathcal{N}(\mathbf{Z}_t|\mathbf{0},\mathbf{I})$ is the regularization term which makes each encoder element uncorrelated and normally distributed. Moreover, we define $P_\phi(\mathbf{X}_t^a|\mathbf{Z}_t)\sim\mathcal{N}(\mathbf{X}_t^a|\mu(\mathbf{Z_t}),\sigma^2(\mathbf{Z}_t))$, $Q_\gamma(\mathbf{Z}_t|\mathbf{X}_t^a)\sim\mathcal{N}(\mathbf{Z_t}|\mu(\mathbf{X}_t^a),\sigma^2(\mathbf{X}_t^a))$ and $P(\mathbf{Z}_t)\sim\mathcal{N}(\mathbf{Z_t}|\mathbf{0},\mathbf{I})$, which are the common choices.

C. DGB: CVAE With Audio-Visual Inputs

Based on the A-DGB framework, we further incorporate vision information using the CVAE scheme. Specifically, DGB extends A-DGB with an auxiliary visual input $\mathbf{X}_t^v$ and an additional Gaussian-like output $d^v(\tau) = exp(-\frac{|\tau-\tilde{\tau}^v|^2}{2\sigma_D^2})$. The entire training criterion modified from (5) is defined as:

$$\begin{aligned}\mathcal{L}^{av} &= \mathbb{E}\left(logP_\gamma(\mathbf{X}_t^a|\mathbf{Z}_t,\mathbf{X}_t^v)\right)\\ &\quad - KL\left(Q_\phi(\mathbf{Z}_t|\mathbf{X}_t^a,\mathbf{X}_t^v)|P(\mathbf{Z}_t)\right).\end{aligned} \tag{6}$$

The encoder and decoder are designed with four stacked convolutional blocks. Each block consists of a 1-D convolutional layer with a single stride and a filter size of 4, a batch normalization layer, and an ELU activation. Smoothed L1 loss is used for reconstruction. By using the re-parameterization trick $\mathbf{Z}_t = \mu(\mathbf{X}_t^a,\mathbf{X}_t^v) + \sigma(\mathbf{X}_t^a,\mathbf{X}_t^v)\odot\xi$ with $\xi\sim\mathcal{N}(\xi|\mathbf{0},\mathbf{I})$, sampling $\mathbf{Z}_t$ from $Q_\phi(\mathbf{Z}_t|\mathbf{X}_t^a,\mathbf{X}_t^v)$ is equal to sampling ξ from $\mathcal{N}(\xi|\mathbf{0},\mathbf{I})$, which is independent of the decoder parameters γ.

The networks are optimized with (5) or (6) with audio or audio-visual signals. The resulting Gaussian-like correlation functions are then fed into the back-end beamformer.

D. Beamforming

The SRP-PHAT [8] is a grid-searching method that estimates the sound at the point with the maximum objective function of SRP. It consolidates all the GCC-PHATs obtained by the array and thus is less sensitive to the TDoA error at each individual microphone pair. For our proposed DGB, instead of using GCC-PHAT, the objective function at an arbitrary grid position $\mathbf{p}$ is built with the A-DGB outputs:

$$G^a(\mathbf{p}) = \frac{1}{M}\sum_m d^m(\tau^m(\mathbf{p})), \tag{7}$$

where $\tau^m(\mathbf{p})$ is the ideal TDoA of the m-indexed pair at $\mathbf{p}$:

$$\tau^m(\mathbf{p}) = \frac{||\mathbf{p}-\mathbf{p}_{n_1}||_2 - ||\mathbf{p}-\mathbf{p}_{n_2}||_2}{c}f_a, \tag{8}$$

where c is the speed of sound in the air, f_a is the audio sampling frequency, and $||\cdot||_2$ is the Euclidean distance.

For our proposed audio-visual DGB, we extend the designed SRP-PHAT objective function (7) by modeling the camera as a virtual microphone pair. Given the camera's known 3D position $\mathbf{p}_c$, the "microphones" are placed at:

$$(\mathbf{p}_{c_1},\mathbf{p}_{c_2}) = \mathbf{p}_c \pm \varepsilon, \tag{9}$$

where ε is a user-defined 3D translation vector.

Then, SRP-PHAT is revised by incorporating the auxiliary Gaussian-like correlation function derived from vision:

$$G^{av}(\mathbf{p}) = \frac{M}{M+1}\sum_m\left(d^m(\tau^m(\mathbf{p}))\right) + \frac{1}{M+1}d^v\left(\tau(\mathbf{p})\right), \tag{10}$$

where $\tau^v(\mathbf{p})$ is the ideal TDoA of the camera, computed by replacing $(\mathbf{p}_{n_1},\mathbf{p}_{n_2})$ with $(\mathbf{p}_{c_1},\mathbf{p}_{c_2})$ in (8).

Finally, the speaker DoA is estimated at the location with the maximum SRP-PHAT value, calculated as:

$$\hat{\theta} = \angle\text{argmax}_{\mathbf{p}\in\mathbf{P}}\, G^{av}(\mathbf{p}), \tag{11}$$

where $\mathbf{P}$ indicates the grid points and '$\angle$' extracts the DoA w.r.t. the microphone array.

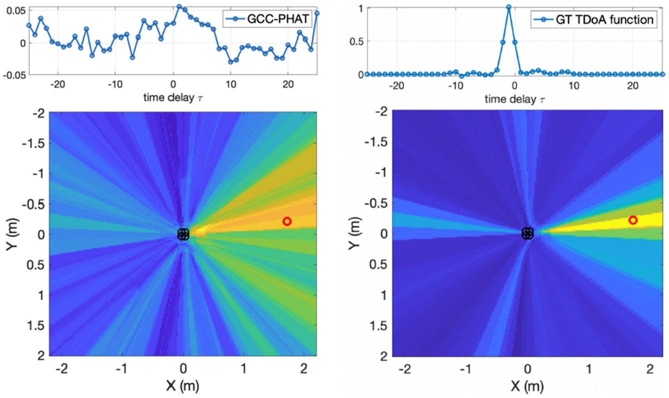
III. EXPERIMENTS

A. A Dataset and Baselines

We conduct experiments on the publicly available SSLR dataset [12] which provides synchronized audio and visual signals, annotated sensor and speaker locations. The dataset was collected by a Softbank's robot Pepper with four microphones mounted on the top of its head, forming a rectangle of 5.8 × 6.9 cm. The audio signals were sampled at 48 kHz. Images were recorded by the Pepper's front camera at 30 fps. We compare with frame-level localization methods, namely Face-MLP, SRP-PHAT [8], GCC-MLP [12], DeepGCC [15], and AV-MLP [32].

TABLE II
EXPERIMENTAL RESULTS OF NOISE ROBUSTNESS ON TEST DATA CORRUPTED AT DIFFERENT SNR LEVELS

		20 dB		10 dB		0 dB		-10 dB		-20 dB	
	Ref	MAE(°)	ACC (%)	MAE(°)	ACC (%)	MAE(°)	ACC (%)	MAE(°)	ACC (%)	MAE(°)	ACC (%)
Audio	SRP-PHAT [8]	3.58	84.6	4.73	74.1	9.11	54.8	27.92	26.1	91.8	2.4
	GCC-MLP [31]	5.41	94.0	7.21	89.0	15.65	75.9	24.87	48.6	43.80	4.4
	deepGCC [15]	5.23	88.5	6.81	86.0	13.30	69.4	20.24	34.0	45.88	1.6
	A-DGB (ours)	4.51	94.9	5.07	91.9	12.32	78.6	13.60	54.6	31.90	5.1
Audio-visual	AVMLP [20]	2.69	95.8	3.27	91.9	5.29	82.5	9.56	57.9	19.59	24.0
	DGB (ours)	**2.23**	**96.7**	**2.24**	**95.2**	**3.97**	**85.4**	**8.29**	**57.4**	**16.48**	**24.0**



(a) (b)

Fig. 2. From top to bottom: (a) the conventional GCC-PHAT and the SRP-PHAT [8] map; (b) our DGB-leanrt correlation function and the resulting acoustic map (Blue/yellow correspond to the lower/higher probability of including a sound source. Red circle indicates the real sound location.)

B. Implementation Details

We re-implement and train the baseline methods with the same parameter settings as our proposals for fair comparison. Mean Absolute Error (MAE) and ACCuracy (ACC) are used. In particular, MAE (°) calculates the average difference between the real and estimated DoA. ACC provides the ratio of frames with correct estimates (the tolerance equals 5°).

We use the RetinaFace detector [33] for face detection. Given the 4-channel array, We adopt all possible pairs and the total microphone pair number $M = 6$. The maximum allowable time-delay length L equals 51. The standard deviation σ_D of the VAE output equals 1 (4). The 3D translation vector equals $\varepsilon = (0, 5, 0)\ cm$ (9). The diagonal elements of the face detection covariance matrix (3) are proportional to the detection width and height (i.e., $w_d = \frac{w*L}{W}$ and $h_d = \frac{h*L}{H}$ with W and H the image width and height). We select Adam [34] as the optimizer, where all models are trained for 40 epochs with a batch size of 256 samples and a learning rate of 0.001.

C. Results

Table I lists the experimental results. The worst performance of Face-MLP (MAE $= 16.85°$ and ACC $= 32.2\%$) verifies the need for input of the audio signal. In addition, we can see that the proposed DGB achieves competitive performance. Specifically, A-DGB results in a MAE of 3.12° and a highly improved ACC of 96.4% given the audio-only inputs. With visual contributions, the better performance is obtained by DGB with the resulting MAE of 1.84° and ACC of 98.0%.

Acoustic signals are susceptible to noise, while an ideal SSL method should also perform well in low SNR conditions. To this end, despite the SSLR dataset already contains robotic ego-noise and slight environmental noise, we train all the models at each epoch where 50% of the original data is corrupted by white Gaussian noise with equal partition at SNRs of 20, 10, 0, -10, and -20 dB, respectively. The evaluation results at different SNR levels are given in Table II where A-DGB has a greater improvement than SRP-PHAT, especially in decreasing SNR. This is because SRP-PHAT is a parametric estimation method that depends on instant GCC-PHAT inputs, while DGB is trained on large-scale annotated data. Comparing with GCC-MLP [12], DGB also superiors, because it incorporates VAE to reconstruct ideal correlations given the noisy GCC-PHAT inputs, thus has better denoising capability. For deepGCC [15], it uses Auto Encoder (AE) to separately encode GCC-PHAT in each individual microphone pair. While our proposed DGB considers the inter-channel relationship by entirely encoding all the observed GCC-PHATs, thus leads to better results. Improvement can also be observed when comparing DGB with AVMLP.

Fig. 2 provides the visualization. The top rows indicate (a) the original GCC-PHAT and (b) the ideal correlation function learned by DGB (for brevity, we only draw the single-channel case). The bottom rows correspond to the resulting acoustic map where yellow/blue means the higher/lower probability of including a sounding object. The sound ground truth location is marked with a red circle. From the figure, we can see that our proposed DGB can derive a much smoother correlation function than the original GCC-PHAT, thus performing better than the original SRP-PHAT algorithm.

IV. CONCLUSION

In this paper, we propose DGB, a novel localization method with audio-visual information incorporated in a VAE framework. Different from the existing DL methods directly map acoustic features to sound locations, given the received noisy GCC-PHATs, DGB reconstructs ideal correlation functions received by each microphone pair. In particular, the monocular camera is modeled as a virtual microphone pair, where the detected face bounding box contributes to an auxiliary correlation function to the back-end beamformer, which locates the sound at the maximum value. Experiments validate the superiority of our proposals over SP- and DL-based approaches.

REFERENCES

[1] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1404–1415, 2020.

[2] R. Tao, Z. Pan, R. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 3927–3935.

[3] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2013.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[5] J. P. Dmochowski and J. Benesty, "Steered beamforming approaches for acoustic source localization," in *Speech Processing in Modern Communication*, Berlin, Germany: Springer, 2010, pp. 307–337.

[6] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[7] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 125–128.

[8] M. Omologo, P. Svaizer, and R. De Mori, "Acoustic transduction," in *Spoken Dialogue with Computer*, New York, NY, USA: Academic, 1998, ch. 2, pp. 1–46.

[9] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Proc. Microphone Arrays*, 2001, pp. 157–180.

[10] D. Antoine, *Acoustic space mapping: A machine learning approach to sound source separation and localization.* Diss. Grenoble, 2013.

[11] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, Mar. 2019.

[12] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 74–79.

[13] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 2814–2818.

[14] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," *Sensors*, vol. 18, no. 10, 2018, Art. no. 3418.

[15] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Acoustic source localization with deep generalized cross correlations," *Signal Process.*, vol. 187, pp. 108–130, 2021.

[16] M. J. Bianco, S. Gannot, E. Fernandez-Grande, and P. Gerstoft, "Semi-supervised source localization in reverberant environments with deep generative modeling," *IEEE Access*, vol. 9, pp. 84956–84970, 2021.

[17] J. Pu, Y. Panagakis, S. Petridis, J. Shen, and M. Pantic, "Blind audio-visual localization and separation via low-rank and sparsity," *IEEE Trans. Cybern.*, vol. 50, no. 5, pp. 2288–2301, May 2020.

[18] V. Sanguineti, P. Morerio, A. Del Bue, and V. Murino, "Audio-visual localization by synthetic acoustic image generation," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, vol. 1, pp. 1138–1145.

[19] Y. Masuyama, Y. Bando, K. Yatabe, Y. Sasaki, M. Onishi, and Y. Oikawa, "Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 4848–4854.

[20] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Audio-visual tracking of concurrent speakers," *IEEE Trans. Multimedia*, vol. 24, pp. 942–954, 2022.

[21] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 288–292, May 1997.

[22] P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1997, vol. 1, pp. 231–234.

[23] D. Blatt and A. O. Hero, "Energy-based sensor network source localization via projection onto convex sets," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3614–3619, Sep. 2006.

[24] K. C. Ho and M. Sun, "Passive source localization using time differences of arrival and gain ratios of arrival," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 464–477, Feb. 2008.

[25] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[26] K. Youssef, S. Argentieri, and J.-L. Zarader, "A binaural sound source localization method using auditive cues and vision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 217–220.

[27] D. Florencio, C. Zhang, and Z. Zhang, "Why does PHAT work well in low noise reverberant environments?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 2565–2568.

[28] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 20–39.

[29] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.

[30] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.

[31] W. He, P. Motlicek, and J.-M. Odobez, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 770–774.

[32] X. Qian, M. Madhavi, Z. Pan, J. Wang, and H. Li, "Multi-target DOA estimation with an audio-visual fusion mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 2814–2818.

[33] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5202–5211.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, 2015, pp. 1412–1420.