LEARNING DYNAMIC STREAM WEIGHTS FOR LINEAR DYNAMICAL SYSTEMS USING NATURAL EVOLUTION STRATEGIES

Christopher Schymura and Dorothea Kolossa

Institute of Communication Acoustics, Faculty of Electrical Engineering and Information Technology, Ruhr University Bochum, Germany

ABSTRACT

Multimodal data fusion is an important aspect of many object localization and tracking frameworks that rely on sensory observations from different sources. A prominent example is audiovisual speaker localization, where the incorporation of visual information has shown to benefit overall performance, especially in adverse acoustic conditions. Recently, the notion of dynamic stream weights as an efficient data fusion technique has been introduced into this field. Originally proposed in the context of audiovisual automatic speech recognition, dynamic stream weights allow for effective sensory-level data fusion on a per-frame basis, if reliability measures for the individual sensory streams are available. This study proposes a learning framework for dynamic stream weights based on natural evolution strategies, which does not require the explicit computation of oracle information. An experimental evaluation based on recorded audiovisual sequences shows that the proposed approach outperforms conventional methods based on supervised training in terms of localization performance.

***Index Terms*—** data fusion, dynamic stream weights, natural evolution strategies, audiovisual speaker localization, Kalman filter

1. INTRODUCTION

Multimodal signal processing is a widely investigated topic in many different application areas, ranging from autonomous driving [1] and smart home applications [2] to object localization and tracking [3]. A specific problem domain of the latter is audiovisual speaker localization. Various studies have shown that an incorporation of the visual modality leads to improved performance over conventional acoustic localization frameworks, especially in noisy and highly reverberant conditions [4, 5]. The use of acoustic and visual modalities has gained similar success in related fields like human-robot interaction [6], speaker diarization [7], speaker identification [8] and automatic speech recognition (ASR) [9–12].

An essential aspect of all multi-modal processing frameworks is the appropriate fusion of individual sensory modalities [13]. This becomes particularly important in highly dynamic scenarios, where the reliability of each sensory input is changing over time. In the context of audiovisual speaker localization, this might occur if the acoustic signals are corrupted by non-stationary background noise or if speakers can not be detected visually, e.g. because they are looking away from the camera. Several approaches to cope with this challenge have been proposed in past studies.

The work introduced in [14] utilizes a Kalman filter (KF)-based framework with a joint audiovisual observation vector to localize and track speakers during recorded seminars. However, it did not incorporate any reliability measures into the processing framework. Instead, data fusion is handled implicitly during the KFs recursive update step. A different approach was proposed in [15], where a specific variant of the particle filter (PF) was utilized to localize and track speakers in a realistic living room scenario. This framework allowed to explicitly control the individual contribution of acoustic and visual input signals via exponential weighting parameters. These parameters were fixed and determined via a grid search over a set of recorded audiovisual scenarios. Additionally, many other approaches to tackle the challenge of audiovisual fusion have been proposed, cf. [7, 16].

Recently, the notion of dynamic stream weights (DSWs) was introduced in the context of audiovisual speaker localization [17]. DSWs were initially proposed for audiovisual ASR [9, 10] as a means to weight acoustic and visual observations on a per-frame basis. Compared to conventional data fusion techniques, this has the advantage that the contribution of each modality can be controlled precisely without significant delay. If additional information about the reliability of sensors is available, this can yield significant improvements in performance over the use of unimodal or naïve fusion techniques [9, 18]. The framework proposed in [17] showed that similar improvements can be achieved for audiovisual speaker localization by incorporating DSWs into a recursive Bayesian state estimator for linear dynamical systems (LDSs). Additionally, a method to obtain oracle dynamic stream weights (ODSWs) from annotated audiovisual sequences, similar to the approach proposed for ASR in [9], was introduced in this study. The ability to compute ODSWs allows DSW estimation models to be learned from annotated training data, which can subsequently be deployed to unseen test scenarios. This has been extensively studied for ASR, where different acoustic and visual features have been used to predict DSWs at each frame [10, 18].

The present study investigates the problem of DSW estimation for LDSs in the context of audiovisual speaker localization and tracking. In particular, an inherent problem of DSW estimation based on oracle information is addressed in this work: the analytic computation of ODSWs requires a prior to be imposed on the DSW distribution, cf. [9, 17], which artificially restricts the representational flexibility of the oracle values. Furthermore, additional hyperparameters that have to be tuned via, e.g., a grid search are introduced into the learning framework.

This work proposes a new learning scheme for DSW prediction models using natural evolution strategies (NES) [19]. NES comprise a class of black-box optimization techniques, that have recently shown success in reinforcement learning (RL) based on deep neural networks (DNNs) [20]. Compared to conventional gradient-descent-based optimizers, NES are computationally more demanding, but do not impose any major restrictions on the underlying cost function (e.g. differentiability). This allows for the derivation of flexible optimization schemes for learning DNN-based DSW prediction models without the need to explicitly compute oracle information.

2. SYSTEM OVERVIEW

This section reviews the audiovisual speaker localization and tracking system introduced in [17], where DSWs were utilized in a framework based on LDSs. Additionally, the analytic computation of ODSWs, also proposed in [17], is briefly described. A basic estimator for DSWs inspired by the work in [9, 18] will be introduced at the end of this section.

2.1. State estimation incorporating dynamic stream weights

In the context of speaker localization and tracking applications, the state of a system that models this process usually refers to the speaker position and additional dynamic properties like velocity and acceleration. Throughout this work, Langevin dynamics [21] are used to model a speaker's motion. This is incorporated into a bimodal LDS model

$$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} + \boldsymbol{v}_k, \tag{1}$$

$$\boldsymbol{y}_{\mathrm{A},k} = \boldsymbol{C}_\mathrm{A}\boldsymbol{x}_k + \boldsymbol{n}_{\mathrm{A},k}, \tag{2}$$

$$\boldsymbol{y}_{\mathrm{V},k} = \boldsymbol{C}_\mathrm{V}\boldsymbol{x}_k + \boldsymbol{n}_{\mathrm{V},k}, \tag{3}$$

where $\boldsymbol{x}_k$ denotes the state vector at discrete time step k, while $\boldsymbol{y}_{\mathrm{A},k}$ and $\boldsymbol{y}_{\mathrm{V},k}$ are conditionally independent acoustic and visual observations. In the context of audiovisual speaker localization, the state usually corresponds to the latent speaker direction-of-arrival (DoA) or Cartesian position, whereas the observations are position-related features. The system dynamics are modeled via the state-transition matrix $\boldsymbol{A}$, subject to zero-mean Gaussian noise $\boldsymbol{v}_k \sim \mathcal{N}(\boldsymbol{0},\, \boldsymbol{Q})$ with covariance matrix $\boldsymbol{Q}$. The acoustic and visual state-to-observation transformations are denoted as $\boldsymbol{C}_\mathrm{A}$ and $\boldsymbol{C}_\mathrm{V}$, respectively. Both observations are assumed to be affected independently by zero-mean Gaussian noise terms $\boldsymbol{n}_{\mathrm{A},k} \sim \mathcal{N}(\boldsymbol{0},\, \boldsymbol{R}_\mathrm{A})$ and $\boldsymbol{n}_{\mathrm{V},k} \sim \mathcal{N}(\boldsymbol{0},\, \boldsymbol{R}_\mathrm{V})$ with covariance matrices $\boldsymbol{R}_\mathrm{A}$ and $\boldsymbol{R}_\mathrm{V}$.

Given sequences of conditionally independent acoustic and visual observations up to time-step k and introducing scalar DSWs $\lambda_k \in [0,\, 1]$, the joint probability density function (PDF) of the LDS introduced in Eqs. (1)–(3) can be expressed as

$$p(\boldsymbol{x}_0,\, \ldots,\, \boldsymbol{x}_k,\, \boldsymbol{y}_{\mathrm{A},1},\, \ldots,\, \boldsymbol{y}_{\mathrm{A},k},\, \boldsymbol{y}_{\mathrm{V},1},\, \ldots,\, \boldsymbol{y}_{\mathrm{V},k}) \propto$$
$$p(\boldsymbol{x}_0) \prod_{k'=1}^{k} p(\boldsymbol{x}_{k'}|\boldsymbol{x}_{k'-1}) p(\boldsymbol{y}_{\mathrm{A},k'}|\boldsymbol{x}_{k'})^{\lambda_{k'}} p(\boldsymbol{y}_{\mathrm{V},k'}|\boldsymbol{x}_{k'})^{1-\lambda_{k'}} \tag{4}$$

A probabilistic framework to infer the state from audiovisual observations can be derived using Eq. (4), yielding a recursive Bayesian estimation scheme similar to the Kalman filter [22]. For a detailed overview of the resulting inference algorithm, cf. [17].

2.2. Oracle dynamic stream weights

To learn DSW estimation models using supervised methods, ODSWs are required as targets [9]. A means to obtain such oracle information from audiovisual observation sequences with a corresponding ground-truth state sequence was introduced in [17]. The proposed method utilized a Gaussian prior with mean μ_λ and variance σ_λ^2 to derive an analytic solution for computing ODSWs at each time-step:

$$\lambda_k^\star = \mu_\lambda + \sigma_\lambda^2 \log\left\{\frac{p(\boldsymbol{y}_{\mathrm{A},k} \,|\, \boldsymbol{x}_k)}{p(\boldsymbol{y}_{\mathrm{V},k} \,|\, \boldsymbol{x}_k)}\right\} \tag{5}$$

As pointed out in [9, 17, 18], the parameters μ_λ and σ_λ^2 can be appropriately determined via cross-validation.

2.3. Reliability measures

The estimation of DSWs requires measures that correspond to the individual reliability of each sensory modality at every time step. Finding appropriate measures for acoustic and visual sensors is a promising field of research which provides many opportunities for further investigations. However, the search for novel reliability measures is beyond the focus of this study. Hence, a more conservative approach based on the work presented in [9] is taken here. Essentially, three different reliability measures are used throughout this work: one signal-based and two model-based reliability measures.

The instantaneous estimated a-priori signal-to-noise ratio (SNR) ξ_k is exploited here as a signal-based reliability measure of the acoustic modality. The noise power estimate, required for computing the a-priori SNR is obtained using the method from [23] and the clean speech power is estimated by a Wiener filter [24]. Additionally, the model-based reliability measures encompass the instantaneous acoustic and visual observation log-likelihoods, $l_{\mathrm{A},k} = \log\{p(\boldsymbol{y}_{\mathrm{A},k} \,|\, \boldsymbol{x}_k)\}$ and $l_{\mathrm{V,k}} = \log\{p(\boldsymbol{y}_{\mathrm{V},k} \,|\, \boldsymbol{x}_k)\}$, respectively. These measures reflect the current degree of belief in how far the instantaneous acoustic and visual observations match the expected observations of the model. The adoption of these model-based reliability measures was inspired by the work described in [9], where similar measures based on the instantaneous entropy were used. In this study, the described reliability measures are combined in a joint feature vector $\boldsymbol{z}_k = \begin{bmatrix}\xi_k & l_{\mathrm{A},k} & l_{\mathrm{V},k}\end{bmatrix}^\mathrm{T}$, which will be serving as input to a suitable mapping function.

It should be noted that the appropriate choice of reliability measures is an important aspect of DSW estimation. The set of measures used in this study has been chosen empirically based on the findings that have been reported in the context of ASR [9, 10, 18]. However, finding suitable measures through e.g. feature selection or even training DSW estimation functions end-to-end is beyond the scope of this paper and will be addressed in future work.

2.4. Mapping functions for dynamic stream weight estimation

After obtaining ODSWs and the corresponding reliability measures as described in Sections 2.2 and 2.3, it is possible to learn appropriate mapping functions $\hat{\lambda}_k = h(\boldsymbol{z}_k \,|\, \boldsymbol{w})$ with parameters $\boldsymbol{w}$ to estimate DSWs. In this work, two different mapping functions will be investigated using conventional supervised training as a baseline: the logistic function, which has already been used for DSW estimation in the context of ASR [9, 18] and a feed-forward DNN, supporting a broader class of nonlinear mappings from reliability measures to DSWs. Supervised training on the basis of ODSWs is performed using standard stochastic gradient descent (SGD) for both methods, where Eq. (5) is used to provide target values at each time-step. This will be compared against NES-based training, described in detail in the next section, where the explicit computation of ODSWs is not required.

3. NATURAL EVOLUTION STRATEGIES

Evolution strategies (ES) are a class of algorithms to solve black box optimization problems [25], which do not require problem specific knowledge on the objective function, except for the ability to evaluate the "fitness" of this function for specific parameter settings. The main concept of ES is to apply a heuristic search procedure to find a parameter vector that maximizes fitness. The implementation of these search heuristics is loosely inspired by natural evolution: At each iteration, a population of candidate solutions is perturbed and

evaluated, resulting in the best performing candidates to be selected for the next iteration. This procedure is iterated until an optimum is reached. A very prominent and widely used ES algorithm is the covariance matrix adaptation evolution strategy (CMA-ES) [26], which represents the population of candidate solutions as a multivariate Gaussian distribution with full covariance matrix. This method has been successfully applied in various applications with small dimensionality. However, to cope with high-dimensional problems (e.g. training DNNs), different methods have to be taken into account.

3.1. Method overview

The present study utilizes NESs, which are a special class of evolution strategies that iteratively update their search distribution via the natural gradient [19]. Let $p(\boldsymbol{w} \,|\, \theta)$ and $f(\boldsymbol{w})$ denote an arbitrary search distribution with parameters θ and a corresponding fitness function evaluated for parameter vector $\boldsymbol{w}$, then the expected fitness under the search distribution can be expressed as

$$J(\theta) = \mathbb{E}_\theta\{f(\boldsymbol{w})\} = \int f(\boldsymbol{w}) p(\boldsymbol{w} \,|\, \theta)\, \mathrm{d}\boldsymbol{w}. \tag{6}$$

An estimate of the search gradient can be obtained from N samples $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N$ using the so-called log-likelihood trick, which yields

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{w}_n) \nabla_\theta \log\{p(\boldsymbol{w}_n \,|\, \theta)\}. \tag{7}$$

For a detailed derivation of Eq. (7), cf. [19]. The approximated search gradient can subsequently be used to iteratively update the search distribution via gradient ascent. Instead of optimizing Eq. (6) based on the plain search gradient directly, NESs utilize the natural gradient [27] to increase the speed of convergence [19]. If a multivariate Gaussian distribution with diagonal covariance matrix is chosen as the search distribution, a very effective NES optimization algorithm applicable for large problem dimensions can be obtained. This specific implementation of NESs is termed separable natural evolution strategies (SNES), which has shown to be well-suited for evolutionary optimization of high-dimensional problems [19].

3.2. Implementation

Learning mapping functions for DSW estimation using NESs requires a training dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}_\mathrm{A}, \mathcal{Y}_\mathrm{V}, \mathcal{Z}\}$ containing M individual and conditionally independent acoustic and visual observation sequences $\mathcal{Y}_\mathrm{A} = \{\boldsymbol{y}_{\mathrm{A},1}^{(m)}, \ldots, \boldsymbol{y}_{\mathrm{A},K_m}^{(m)}\}_{m=1}^{M}$ and $\mathcal{Y}_\mathrm{V} = \{\boldsymbol{y}_{\mathrm{V},1}^{(m)}, \ldots, \boldsymbol{y}_{\mathrm{V},K_m}^{(m)}\}_{m=1}^{M}$, with corresponding ground-truth state trajectories $\mathcal{X} = \{\boldsymbol{x}_1^{(m)}, \ldots, \boldsymbol{x}_{K_m}^{(m)}\}_{m=1}^{M}$ and reliability measures $\mathcal{Z} = \{\boldsymbol{z}_1^{(m)}, \ldots, \boldsymbol{z}_{K_m}^{(m)}\}_{m=1}^{M}$. In comparison to the ODSW estimation problem [17], explicit computation of oracle information is not required when using NESs. Instead, a fitness function has to be designed, which reflects the objective that should optimized. For the application of speaker localization and tracking as considered in this work, the averaged negative azimuth localization error

$$f(\boldsymbol{w}) = -\frac{1}{M} \sum_{m=1}^{M} \frac{1}{K_m} \sum_{k=1}^{K_m} \left(\phi_k^{(m)} - \hat{\phi}_k^{(m)}(\boldsymbol{w})\right)^2 \tag{8}$$

is chosen as an appropriate measure of fitness that should be maximized. The ground-truth speaker azimuth at time-step k of the m-th sequence is denoted as $\phi_k^{(m)}$ and the corresponding estimated azimuth $\hat{\phi}_k^{(m)}(\boldsymbol{w})$ represents the estimate obtained using the state estimation framework introduced in Sec. 2.1 with a specific parameterization $\boldsymbol{w}$ of the mapping function. Fitness shaping as proposed in [19, Sec. 3.1] is used in this work to increase robustness of the optimization algorithm in the presence of different scalings of the fitness function. The block diagram depicted in Fig. 1 illustrates the fundamental differences between supervised training using ODSWs and the NES-based training paradigm.

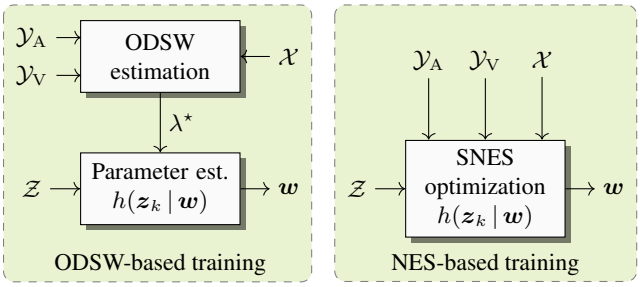

Fig. 1: Block diagrams of the two different training paradigms investigated in this study. Ground-truth state trajectories $\mathcal{X}$, audiovisual observation sequences $\mathcal{Y}_\mathrm{A}$ and $\mathcal{Y}_\mathrm{V}$, as well as the corresponding reliability measures $\mathcal{Z}$ are used as training datasets in both cases. The set of estimated oracle stream weights is denoted as $\lambda^\star$.

4. EVALUATION

A performance evaluation of the proposed methods was conducted in the domain of audiovisual speaker localization. The fundamental question that is addressed in this work is, whether NES-based training can benefit DSW estimation over conventional supervised methods, which require the explicit estimation of ODSWs.

4.1. Dataset

A dataset of audiovisual recordings was collected using the Microsoft Kinect sensor. The sensor was placed in a reverberant office room with a reverberation time of approximately $350\,\mathrm{ms}$. Ten audiovisual sequences of a single moving speaker with a duration of $30\,\mathrm{s}$ each were recorded using the described setup. The video signals were recorded at a frame rate of 20 frames per second (FPS) and the audio signals were captured using the four-channel microphone array of the Kinect at a sampling rate of $16\,\mathrm{kHz}$.

The speed and type of movement, as well as the amount of speech varied between the individual sequences. An example of different conditions captured in the dataset is shown in Fig. 2. All recorded sequences were manually annotated, where the ground-truth locations of the speaker's face were marked at each frame.

4.2. Experimental setup

Acoustic DoA observations were computed at intervals of $50\,\mathrm{ms}$ (matching the frame rate of the video signal) from the recorded microphone array signals using the multiple signal classification (MUSIC) algorithm [28] with direct-path dominance (DPD) test [29] to increase robustness of the acoustic localization in the presence of reverberation. Visual locations of the speaker's face were extracted from the recorded video using the Viola-Jones algorithm [30].
Similar to the previous work proposed in [17], a standard KF serves as a baseline for comparison. Additionally, state estimation based

Fig. 2: Snapshots from a recorded audiovisual sequence used during the experimental evaluation in this work. In the video frame shown on the left, the speaker is facing the camera, which results in a correct detection of the speaker's face (denoted by the yellow rectangle). This is not possible in the frame on the right side, as the speaker is facing to the side and some parts of the face are covered by shadows.

on ODSWs as described in Sec. 2.2 is also evaluated, representing an upper bound on performance. The logistic mapping function for DSW estimation is used without any modifications, similar to the work in [9, 18]. A three-layer feed-forward neural network with rectified linear unit (ReLU) activation functions in the hidden layers and a softmax output activation function is used as the second mapping function. The number of neurons in the hidden layers is determined during nested cross-validation, as described below. Both mapping functions were trained using either conventional supervised SGD with ODSWs as targets and cross-entropy loss, or the SNES algorithm as described in Sec. 3.2.

Performance evaluation was conducted using ten-fold nested cross-validation, where an additional inner cross-validation loop is utilized for additional tuning of hyperparameters. The hyperparameters in this study were mean and variance of the Gaussian prior for estimating ODSWs (cf. Sec. 2.2), as well as the number of neurons in the hidden layers of the neural-network-based mapping function. The azimuth root mean square error (RMSE) was used as a metric to evaluate localization performance. It was computed individually at each frame and subsequently averaged over all sequences.

4.3. Results and discussion

The performance evaluation results are depicted in Fig. 3, where the logistic function ("LF") and a feed-forward neural network ("NN") are used as mapping functions for stream weight estimation, trained with either stochastic gradient descent on ODSWs ("SGD") or using the separable NES algorithm ("SNES"). Both mapping functions achieve lower azimuth error when trained using SNES, compared to SGD-based training. This indicates that the direct utilization of localization error as an objective function via NESs is advantageous over conventional supervised training based on ODSW targets.

The SNES-based methods both outperform the standard KF baseline, whereas SGD training is only able to achieve similar performance to the KF in both cases. It is interesting to note that both the logistic function, as well as the neural network do not differ much in terms of performance when trained using SGD. This indicates that the logistic function already provides adequate representational capabilities when using the three reliability measures proposed in this work. However, the performance of the neural network might be further improved if a larger amount of training data is available. When using the SNES algorithm for training, the neural network slightly outperforms the logistic function, which seems to be due to a better use of the limited training data and a

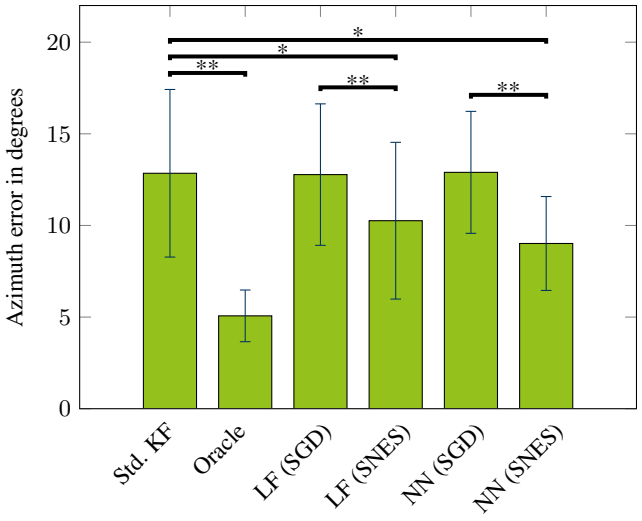

Fig. 3: Results of the experimental comparison between different DSW estimation approaches, averaged over all cross-validation folds. Similar to the work presented in [17], a standard KF with joint observation noise covariance matrix ("Std. KF") serves as a baseline. The asterisks denote important statistically significant differences with $*$ for $p < 0.05$ and $**$ for $p < 0.01$.

reduced risk of overfitting provided by ESs [19, 20]. Despite the extended training time that NES-based methods require over SGD, the improvement in performance indicates that these methods might be interesting candidates for further research in this direction.

The direct use of ODSWs still achieves a significantly lower localization error than all other evaluated methods. As this can be interpreted as an upper bound on achievable performance [9], there is still room for further improvement. However, this outcome additionally confirms the initial results reported in [17] and shows that the incorporation of DSW into the field of LDSs is beneficial to multimodal sensor fusion problems in continuous state spaces.

5. CONCLUSIONS AND OUTLOOK

Focusing attention on different sensory modalities via dynamic stream weights in linear dynamical systems proved itself as an efficient sensor fusion strategy in the context of audiovisual speaker localization, but can also be applied in many other domains. This study has introduced a novel approach to learn mapping functions for dynamic stream weight estimation based on natural evolution strategies. The proposed method does not require the explicit computation of oracle dynamic stream weights, which was required for previously proposed methods. Despite the longer time needed to train systems based on natural evolution strategies, this study has shown that superior performance can be achieved compared to conventional supervised training based on gradient descent.

Future investigations will focus on a thorough analysis of suitable reliability measures. The measures utilized in this study only serve as a starting point for further research. An in-depth analysis of different measures by means of feature selection may yield interesting theoretical insights towards the reliability of audiovisual sensors. Additionally, making the proposed system trainable end-to-end and extending the inference framework to nonlinear dynamical systems are further promising research directions.

6. REFERENCES

[1] T. N. N. Hossein, S. Mita, and H. Long, "Multi-sensor data fusion for autonomous vehicle navigation through adaptive particle filter," in *IEEE Intelligent Vehicles Symposium*, June 2010, pp. 752–759.

[2] H. Medjahed, D. Istrate, J. Boudy, J. Baldinger, and B. Dorizzi, "A pervasive multi-sensor data fusion for smart home healthcare monitoring," in *IEEE International Conference on Fuzzy Systems*, June 2011, pp. 1466–1473.

[3] S. Shahrampour, M. Noshad, J. Ding, and V. Tarokh, "Online Learning for Multimodal Data Fusion With Application to Object Recognition," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 9, pp. 1259–1263, Sept 2018.

[4] A. Ronzhin, A. Ronzhin, and V. Budkov, "Audiovisual speaker localization in medium smart meeting room," in *Proc. of the International Conference on Information, Communications Signal Processing*, December 2011, pp. 1–5.

[5] G. Monaci, "Towards real-time audiovisual speaker localization," in *19th European Signal Processing Conference*, August 2011, pp. 1055–1059.

[6] B. Chen, M. Meguro, and M. Kaneko, "Probabilistic integration of audiovisual information to localize sound source in human-robot interaction," in *Proc. of the 12th IEEE International Workshop on Robot and Human Interactive Communication*, 2003.

[7] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, 2018.

[8] L. Schönherr, D. Orth, M. Heckmann, and D. Kolossa, "Environmentally robust audio-visual speaker identification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016.

[9] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning Dynamic Stream Weights For Coupled-HMM-Based Audio-Visual Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, May 2015.

[10] S. Gergen, S. Zeiler, A. H. Abdelaziz, R. M. Nickel, and D. Kolossa, "Dynamic stream weighting for turbo-decoding-based audiovisual ASR," in *17th Annual Conference of the International Speech Communication Association*, 2016.

[11] J. Freiwald, M. Karbasi, S. Zeiler, J. Melchior, V. Kompella, L. Wiskott, and D. Kolossa, "Utilizing slow feature analysis for lipreading," in *Speech Communication; 13. ITG Symposium*, October 2018.

[12] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2018, pp. 6548–6552.

[13] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, September 2015.

[14] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2005, pp. 118–121.

[15] S. Gerlach, S. Goetze, and S. Doclo, "2D audio-visual localization in home environments using a particle filter," in *Speech Communication; 10. ITG Symposium*, September 2012, pp. 1–4.

[16] R. Yan, T. Rodemann, and B. Wrede, "Computational Audio-visual Scene Analysis in Online Adaptation of Audio-Motor Maps," *IEEE Transactions on Autonomous Mental Development*, vol. 5, no. 4, pp. 273–287, December 2013.

[17] C. Schymura, T. Isenberg, and D. Kolossa, "Extending Linear Dynamical Systems with Dynamic Stream Weights for Audio-visual Speaker Localization," in *IEEE International Workshop on Acoustic Signal Enhancement*, September 2018, pp. 515–519.

[18] H. Meutzner, N. Ma, R. M. Nickel, C. Schymura, and D. Kolossa, "Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[19] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, "Natural Evolution Strategies," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 949–980, January 2014.

[20] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, "Evolution strategies as a scalable alternative to reinforcement learning," *arXiv preprint arXiv:1703.03864*, 2017.

[21] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, May 2001, pp. 3021–3024.

[22] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME – Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, March 1960.

[23] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[24] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, April 1980.

[25] I. Rechenberg and M. Eigen, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.*, Frommann-Holzboog, 1973.

[26] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evol. Comput.*, vol. 9, no. 2, pp. 159–195, June 2001.

[27] S. I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, February 1998.

[28] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, 1986.

[29] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, 2014.

[30] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.