A Novel Proximal Policy Optimization Approach for Filter Design

Dongdong Fan[1], Shuai Ding[1,2], Haotian Zhang[2], Weihao Zhang[4], Qingsong Jia[2], Xu Han[2], Hao Tang[2], Zhaojun Zhu[2], and Yuliang Zhou[3]

[1]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
Shenzhen 518110, China
dond_fan@std.uestc.edu.cn

[2]Institute of Applied Physics, University of Electronic Science and Technology of China
Chengdu 610054, China
uestcding@uestc.edu.cn

[3]School of Aeronautics and Astronautics, University of Electronic Science and Technology of China
Chengdu 610054, China

[4]School of Materials and Energy, University of Electronic Science and Technology of China
Chengdu 610054, P.R. China

***Abstract* –** This paper proposes a proximal policy optimization (PPO) algorithm for coupling matrix synthesis of microwave filters. With the improvement of filter design requirement, the limitations of traditional methods such as limited applicability are becoming more and more obvious. In order to improve the filter synthesis efficiency, this paper constructs a reinforcement learning algorithm based on Actor-Critic network architecture, and designs a unique filter coupling matrix synthesis reward function and action function, which can solve combinatorial optimization problems stably.

***Index Terms* –** bandpass filters (BPF), coupling matrix synthesis, Proximal Policy Optimization (PPO).

I. INTRODUCTION

With the development of wireless communication technologies such as 5G or post-5G, the requirements for the integration and design efficiency of passive microwave devices are increasing, among which filters are the most important ones since they can select specific frequencies. Filter design involves multiple steps and several factors, such as insertion loss, bandwidth, working frequency, out-of-band suppression, physical size, power capacity and stability [1].

Automation of filter design has long been pursued to enhance design efficiency [2]. In recent years, a rising number of artificial intelligent methods have been incorporated in the filter design process. Among them, optimization is a common method in the design process based on electromagnetic simulation. Optimization aims to transform the design specification into a suitable objective function, and then obtain the parameters that meet the final design requirements through an optimization algorithm. For example, rapid simulation and optimization of microwave component models based on functional substitution modeling technology can enable advanced circuit design or computer-aided tuning of microwave components [3], The coupling matrix algorithm based on neural network can realize filter synthesis and fine tuning [4–7], and the adaptive synthesis of resonant-coupled filters can be realized based on particle swarm optimization [8, 9] and spatial mapping technology [10, 11].

In this paper, we propose to solve the filter synthesis problem by applying a proximal policy optimization (PPO) algorithm based on deep reinforcement learning. We construct a neural network model based on the Actor-Critic architecture and design specific reward function and action function to synthesize the filter coupling matrix. The novelty and main contributions of this paper are as follows: (1) to the best of our knowledge, this is the first work to present a complete PPO framework and apply it to the synthesis of filter coupling matrix; and (2) based on extensive experiments, we design a model structure that can solve this problem and achieve satisfactory results.

II. METHODOLOGY: PPO ALGORITHM

A. Framework

A PPO algorithm is a reinforcement learning algorithm proposed by OpenAI in 2017 [12]. It is considered a state-of-the-art method in the field of reinforcement learning and is one of the most widely applicable

Submitted On: December 19, 2023
Accepted On: June 21, 2024

https://doi.org/10.13052/2024.ACES.J.390502

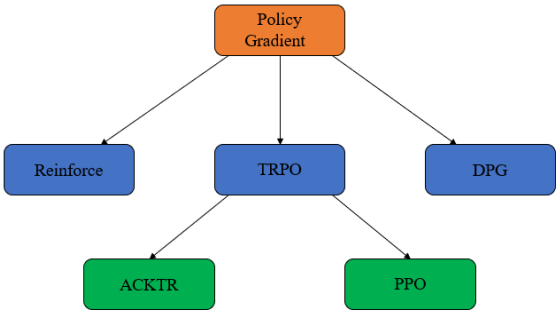


Fig. 1. PPO evolution process.

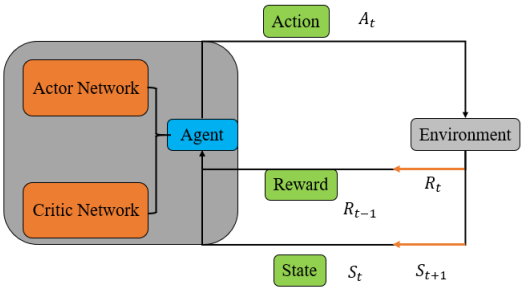


Fig. 2. Actor-Critic architecture.

algorithms in the field. Because it is simple to implement and exhibits stable performance, a PPO algorithm can handle both discrete or continuous action spaces and conduct large-scale training. It has received widespread attention in recent years due to these advantages, and its evolution is shown in Fig. 1.

The core idea of a PPO algorithm is to use PPO to train the agent. PPO is a kind of policy gradient reinforcement learning algorithm that optimizes the policy by maximizing the expected return.The core of a PPO algorithm is the use of the following policy loss function:

$$L_{CLIP}(\theta) = \hat{E}_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)]. \quad (1)$$

Where we have the following definitions.

$r_t(\theta) = (\pi_\theta\ (a|s))/(\pi_{\theta old}\ (a|s))$ is the policy update ratio. The larger the $r_t(\theta)$, the higher the probability of taking action a under state s by the current policy, and the larger the update ratio relative to the old policy.

$\hat{A}_t = Q_{\pi_\theta old}\ (s,a) - V_{\pi\theta old}(s)$ is the advantage function, which represents the difference between the value of the current state and action and the average value, which is used to calculate the clipping range in the proximal ratio clipping loss. The larger the value of the advantage function, the better the current state and action, and they should obtain a larger reward.

ε is a hyper-parameter that controls the clipping range.

$clip(x,a,b)$ is a clipping function, which means that x is restricted to the interval [a,b].

$\hat{E}_t$ represents the expected experience over time steps.

To summarize: the proximal ratio clipping loss consists of two parts, and we chose the smaller one. This can ensure that the policy update does not deviate too much from the original policy, thus achieving stable and efficient training results.

The Actor-Critic network, and their basic architecture are shown in Fig. 2. The Actor network is responsible for outputting policies, i.e., the probability distribution of action selection at each state; the Critic network is responsible for estimating the state value function, i.e., the expected cumulative reward at each state. The core idea of the PPO algorithm is to limit the magnitude of policy updates to ensure that the policy does not deviate too far, thereby improving the stability and efficiency of learning.

B. Coupling matrix synthesis based on PPO

By modeling the comprehensive process of the coupling matrix as a deep reinforcement learning problem, a deep neural network model is trained by taking the performance index of the filter (such as bandwidth and return loss) as the state, the adjustment of the coupling coefficient in the coupling matrix by the agent as the action, and the change of the performance index when the coupling coefficient is adjusted as the reward. The method consists of the following modules.

State and action space: The state space refers to the set S of possible states in the coupling matrix synthesis problem, expressed as follows $S = \{s_1, s_2, , s_n\}$. The action space refers to the set A of all of the possible actions that the agent can take, expressed as follows $A = \{a_1, a_2, , a_m\}$. In this method, the agent uses discrete actions to add or subtract the elements of the coupling matrix with a fixed step length within a certain range to achieve the change of the state.

State transition: In reinforcement learning, a state transition is the agent learning by interacting with the environment, observing the current state, and then acting on its own strategy and receiving a reward or punishment from the environment. It then moves to a new state, and this process is called a state transition. The state transition function is usually expressed as:

$$s' = f(s,a), \quad (2)$$

where s is the current state, a is the action taken by the agent, and $s\prime$ is the new state transferred to by the agent.

Reward function: The reward function is used to evaluate the value of each state and action and is denoted as $R(s_t, a_t, s_{(t+1)})$. In this paper, a special reward function is proposed for coupling matrix synthesis that consists of two parts: target difference reward $R_{S_{11max}}$ and distance reduction reward $R_{S_{reduce}}$. The target difference

reward refers to the absolute difference between the maximum return loss and target return loss, and the absolute difference between the minimum out-of-band return loss and target return loss in the coupling matrix synthesis process. The target difference reward can be written in the following form:

$$R_{S_{11max}} \propto \frac{1}{S_{11max} - S_{11goal}}. \tag{3}$$

$R_{S_{reduce}}$ sets the reward by measuring the difference between the current S-parameter state and next S parameter state by means of the mean square error. We define Dist so as to construct a set consisting of the values of the S parameters of each frequency of the target state, and the mean-square error of the frequency point values between the two sets is calculated using the following formula. After the action is executed, when the Dist of the next state is greater than that of the current state, it means that the agent is moving away from the target, and the reward is 0. Otherwise, the reward is 1, thereby encouraging the agent to execute actions in the direction in which Dist becomes smaller. The award may be written as follows:

$$Dist = \frac{1}{n}\sum_{i=1}^{n}(S_{11real}(i) - S_{11goal}(i))^2, \tag{4}$$

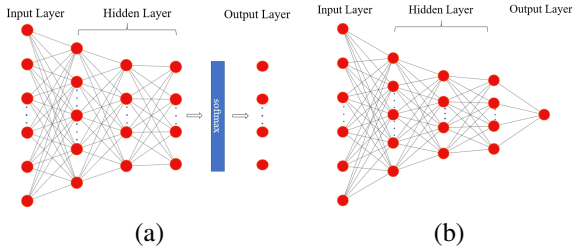
$$R_{S_{reduce}} = \begin{cases} 1, & if newDist > lastDist \\ 0, & if newDist \leq lastDist. \end{cases} \tag{5}$$

Network architecture and training process: The basic structure of the Actor and the Critic network adopts a fully connected neural network and is shown in Fig. 3. The neural network structure in the PPO algorithm consists of an input layer, a hidden layer and an output layer.

Neural network training can be described as an optimization problem, and this optimization algorithm usually needs to calculate the gradient. In the neural network with sigmoid function, the gradient becomes smaller and smaller in the process of backpropagation and gradually approaches zero as the number of layers increases. Gradients approaching zero prevent weights from being updated during training. Such a problem is called the vanishing gradient problem. In fact, when using sigmoid activation functions, the gradient will usually vanish,especially at the beginning of learning [13, 14]. ReLU allows deep neural networks to have no gradient vanishing problem during training[15, 16]. Deep neural networks with ReLU have been proven to be effective for speech recognition[17].

In order to overcome the problem of gradient disappearance during deep neural network training, we use ReLU as the activation function.The ReLU function is expressed as:

$$f(\gamma) = \max(\gamma, 0) = \begin{cases} \gamma, & \text{if } \gamma > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$



(a) (b)

Fig. 3. Actor (a) and Critic (b) network structure.

The gradient of ReLU is:

$$f'(\gamma) = \begin{cases} 1, & \text{if } \gamma > 0 \\ 0, & \text{otherwise} \end{cases}. \tag{7}$$

In the case of a negative input, it will output 0, then the neuron will not be activated. This means that only some neurons are activated at the same time, making the network sparse and thus very efficient for computation.

Step 1: The Actor and Critic networks are constructed by initializing the parameters θ_0 and ω_0.

Step 2: Collect data and store them in experience pool D_0:$D_t = (s_t, a_t, r_t, s_{(t+1)})$, where s_t represents the state at time t, a_t represents the action at time step t, r_t represents the reward at time t, and $s_{(t+1)}$ represents the state at time $t+1$.

Step 3: For each training cycle, we repeat the following steps:

a: Update the experience pool data.
b: The PPO method optimizes the policy function $\theta_k = argmax_\theta L^{CLIP}(\theta_{(k-1)}, \theta)$, where $L^{CLIP}(\theta_{(k-1)}, \theta)$ represents the loss function of the Actor network.
c: We repeat steps a and b until the specified number of training rounds is reached or the convergence condition is reached.

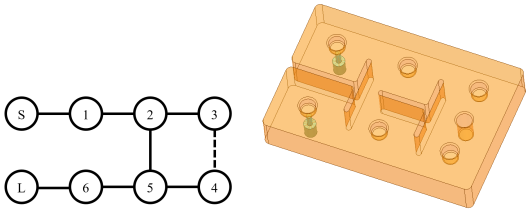
Step 4: Output the optimal policy function and use it to generate the agent's actions $\pi^*(a|s) = arg\max_\theta L^{CLIP}(\pi)$. The optimization process must be limited to ensure that the step size of each update is not too large to avoid excessive updating. The optimal policy function $\pi^*(a|s)$ can be obtained through the Actor network and is used to generate the actions of the agents.

III. DESIGN EXAMPLES

A. Design specification

The sixth-order dielectric waveguide BPF shown in Fig. 4, uses PPO for coupling matrix synthesis.

The design specifications are as follows:
1): Center frequency: f_0=3.0 GHz.
2): Fractional bandwidth: $\Delta f f_0 = 5\%$.
3): Transmission zero: 2780 and 3220 MHz.
4): Number of resonators: $N_R = 6$.
The dielectric constant of the dielectric waveguide filter is 20.5.



(a) (b)

Fig. 4. (a) Sixth order filter topology with symmetric transmission zeros, and (b) 3D model.

B. Concrete realization

During the coupling matrix synthesis process, when the intelligent agent interacts with the environment, the first step involves acquiring the current state. In this case, a sixth-order filter with a center frequency of 3 GHz and a bandwidth of 150 MHz is employed. Based on the symmetry of the coupling matrix, there are a total of eight nonzero values in the current state. The role of the intelligent agent is to modify these eight values by either increasing or decreasing them, with the values along the diagonal of the coupling matrix ranging from [0.5,1.3]. The range for the cross-coupling $m_{2,5}$ is $[-0.5,0.5]$.

According to the symmetry of the coupling matrix, the agent has a total of 10 different actions, which are expressed as follows:

$$A = \{m_{0,1}\pm, m_{1,2}\pm, m_{2,3}\pm, m_{3,4}\pm, m_{2,5}\pm\}, \quad (8)$$

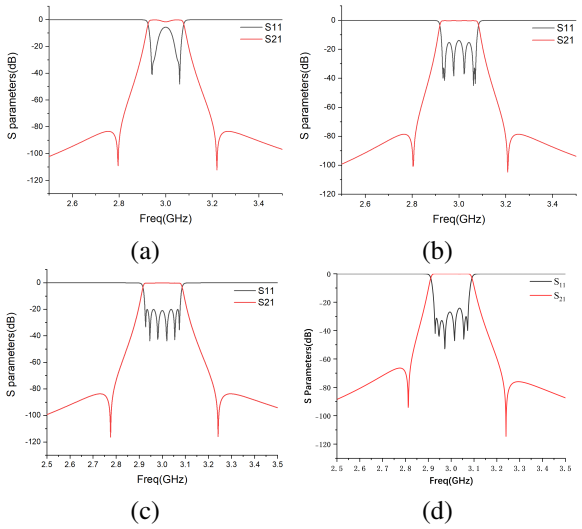
$$m_{(x,x)}\pm = \max\left(\min\left(m_{(x,x)} \pm change, \max M\right), \min M\right), \quad (9)$$

where $m_{(x,x)}$ refers to the currently adjusted matrix element.

After the agent performs an action, it transitions to a new state. In this process, the environment provides the agent with rewards for the action based on the old state and the new state. As described in section II, these rewards are utilized by the agent through the use of a reward function to select actions for coupling matrix synthesis, thus adjusting the direction and objectives for the coupling matrix. During the agent's training process, the PPO network gradually learns the relationship between S-parameters and the adjustment of the coupling matrix. The agent continues to explore in search of better solutions. Upon training completion, the agent is capable of identifying coupling matrices that meet the specified requirements.

C. Design results

In this paper, a PPO algorithm is used to optimize the eighth order coupling matrix for 2 minutes, which is a relatively long value in the optimization process because the PPO algorithm itself has randomness. The comprehensive process is shown in Figs. 5 (a)-(c). The eight nonzero coupling coefficients M of this sixth-order filter are $M = \{M_{0,1} = M_{6,7} = 1.009, M_{1,2} = M_{5,6} = 0.851, M_{2,3} = M_{4,5} = 0.617, M_{3,4} = 0.61, M_{2,5} = -0.025\}$. As can be seen from Fig. 5 (c), the S parameters meet the in-band return loss and insertion loss, and generate two transmission zeros near the specified frequency. In addition, according to the synthesized coupling matrix, a full-wave simulation was performed in simulation software Ansys HFSS, and the simulation result is shown in Fig. 5 (d). The simulation result is basically consistent with the S parameters of the filter synthesized by the coupling matrix.



(a) (b) (c) (d)

Fig. 5. (a)-(c) Coupling matrix synthesis process based on PPO algorithm, and (d) full-wave simulation result.

IV. CONCLUSION

In this paper, a PPO algorithm in deep reinforcement learning is introduced, and an Actor-Critic network for coupling matrix synthesis is constructed and designed with unique action function and reward function. The coupling matrix of a six-order filter is synthesized by using PPO, and the corresponding full-wave simulation is performed after obtaining the coupling matrix. It is proved that the S parameters of the coupling matrix synthesis and the full-wave simulation results corresponding to the coupling matrix are basically consistent. The feasibility and generality of the PPO algorithm are verified. In the proposed PPO comprehensive coupling matrix in this paper, although the synthesis time for complex coupling matrices is relatively long, this algorithm not only synthesizes traditional common coupling matrices but also can synthesize some special coupling matrices. That is, it can synthesize uncommon coupling structures.

ACKNOWLEDGMENT

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant A03019023801088, Sichuan Province Science and Technology Support Program,(No.2022YFS0193) and Fundamental Research Funds for the Central Universities(No.ZYGX 2021YGLH025).

REFERENCES

[1] D. Liang, X. Zhang, B. Yang, and D. Young, "Overview of base station requirements for RF and microwave filters," in *2021 IEEE MTT-S International Microwave Filter Workshop (IMFW)*, pp. 46-49, 2021.

[2] R. J. Cameron, C. M. Kudsia, and R. R. Mansour, *Microwave Filters for Communication Systems: Fundamentals, Design, and Applications*. Hoboken, NJ: John Wiley & Sons, 2018.

[3] F. Feng, C. Zhang, J. Ma, and Q.-J. Zhang, "Parametric modeling of EM behavior of microwave components using combined neural networks and pole-residue-based transfer functions," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 1, pp. 60-77, 2016.

[4] M. Ohira, A. Yamashita, Z. Ma, and X. Wang, "Automated microstrip bandpass filter design using feedforward and inverse models of neural network," in *2018 Asia-Pacific Microwave Conference (APMC)*, pp. 1292-1294, 2018.

[5] M. Ohira, K. Takano, and Z. Ma, "A novel deep-Q-network-based fine-tuning approach for planar bandpass filter design," *IEEE Microwave and Wireless Components Letters*, vol. 31, no. 6, pp. 638-641, 2021.

[6] M. Ohira, A. Yamashita, Z. Ma, and X. Wang, "A novel eigenmode-based neural network for fully automated microstrip bandpass filter design," in *2017 IEEE MTT-S International Microwave Symposium (IMS)*, pp. 1628-1631, 2017.

[7] B. Liu, H. Yang, and M. J. Lancaster, "Global optimization of microwave filters based on a surrogate model-assisted evolutionary algorithm," *IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 6, pp. 1976-1985, 2017.

[8] J. L. Chvez-Hurtado and J. E. Rayas-Snchez, "Polynomial-based surrogate modeling of RF and microwave circuits in frequency domain exploiting the multinomial theorem," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 12, pp. 4371-4381, 2016.

[9] L. Bi, W. Cao, W. Hu, and M. Wu, "Intelligent tuning of microwave cavity filters using granular multi-swarm particle swarm optimization," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 12, pp. 12901-12911, 2021.

[10] S. Koziel, J. Meng, J. W. Bandler, M. H. Bakr, and Q. S. Cheng, "Accelerated microwave design optimization with tuning space mapping," *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 2, pp. 383-394, 2009.

[11] Q. S. Cheng, J. W. Bandler, and S. Koziel, "Space mapping design framework exploiting tuning elements," *IEEE Transactions on Microwave Theory and Techniques*, vol. 58, no. 1, pp. 136-144, 2010.

[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[13] M. A. Nielsen, Neural Networks and Deep Learning, vol. 25. San Francisco, CA: Determination Press, 2015.

[14] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107-116, 1998.

[15] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[16] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Journal of Machine Learning Research*, vol. 15, pp. 315-323, 2011.

[17] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8609-8613, 2013.

Dongdong Fan was born in Shanxi, China,in 1996. He received the B.E. degree in Optoelectronic information science and engineering from the Nanyang Institute of Technology of China, in 2019, where he is currently pursuing the M.E. degree in electronic information engineering with the School of Physics. His current research interests include radio-frequency circuit and filter.

Shuai Ding received the Ph.D. degree in radio physics from the University of Electronic Science and Technology of China (UESTC), Chengdu, in 2013. From 2013 to 2014, he was a Postdoctoral Associate with the cole Polytechnique de Montral, Montral, QC, Canada.

In 2015, he joined UESTC, where he is currently an Associate Professor. He has authored or coauthored over 80 publications in refereed journals and international conferences/symposia. His current research interests include time-reversed electromagnetics and its applications to communication and energy transmission, phased array, analog signal processing, and microwave circuits. He has served as a TPC Member for various conferences and a reviewer for several peer-reviewed periodicals and international conferences/symposia.

Haotian Zhang was born in Henan, China,in 1998. He is received the M.E. degree in physics with the School of Physics from the University of Electronic Science and Technology of China, in 2023. His current research interests include machine learning, antenna arrays, and filter.

Weihao Zhang was born in Handan, China,in 1995. He received theB.E. degree in fundamental science (mathematics and physics) from the University of Electronic Science and Technology of China, in 2018, where he is currently pursuing the Ph.D.degree in Electronic Information Materials and Components with University of Electronic Science and Technology of China, ChengDu, China. His current research interests include integrated magnetic devices and fabrication technologies, filter, filtering antenna, metasurface, antenna array.

Qingsong Jia was born in Sichuan, China,in 1997. He received the B.E. degree in electronic information science and technology from the University of Electronic Science and Technology of China, in 2019, where he is currently pursuing the Ph.D.degree in electromagnetic field and microwave technology with the School of Physics. His current research interests include metasurface, antenna arrays, and the application of radio OAM vortex wave.

Xu Han was born in Sichuan, China, in 1995. He received the B.E. degree in electronic information science and technology from the University of Electronic Science and Technology of China, in 2018, where he is currently pursuing the Ph.D.degree in electromagnetic field and microwave technology with the School of Physics. His current research interests include metasurface, antenna arrays, and phase array.

Hao Tang was born in Hebei, China, in 1998. He received the B.E. degree in Internet of Things Engineering from the Chengdu University of Technology of China,in 2020, where he is currently pursuing the Ph.D. degree in physics with the School of Physics. His current research interests include metasurface and antenna arrays.

Zhaojun Zhu was born in Sichuan, China, in 1978. He received the B.S. degree and the Ph.D.degree in physical electronics from the University of Electronic Science and Technology of China(UESTC), Chengdu, in 2002 and 2007, respectively. Since 2012, he has been an Associate Professor with UESTC. His research interests include the design of microwave and millimeter-wave circuits.

Yuliang Zhou received the B.S. degree in applied physics from the University of Electronic Science and Technology of China, Chengdu, China, in 2012, where he is currently pursuing the Ph.D. degree in communication and information systems. From 2017 to 2018, he was with the Microwave Laboratory, University of Pavia, Pavia, Italy. His current research interests include substrate integrated circuits, leaky-wave antennas, and systems for wireless communication.